Addressing Psychosocial and Lifestyle Risk Factors to Promote Primary Cancer Prevention: an integrated platform to promote behavioural change (IBeCHANGE)

Project Number: 101136840

D3.1 – Analysis of Retrospective Data

Related Work Package	**WP3 – Novel approaches for interaction through a Virtual User Model**
Related Task	**T3.1 Collection and analysis of retrospective data (Lead: ICO; Participants: IEO, SD, SIMAVI, UMFCD)**
Lead Beneficiary	ICO
Contributing Beneficiaries	IEO, SD, SIMAVI, UMFCD
Document version	v1.0
Deliverable type	Document Report
Dissemination level	RE/SEN
Due date	31/05/2025
Delivery date	31/07/2025
Authors	Mireia Obón-Santacana (ICO); Elisabet Guinó (ICO); Noemie Travier (ICO); Aline Machiavelli (SD)
Contributors	Anna García Serra (ICO); Ricardo Pietrobon (SD); Marianna Masiero (IEO); Chiara Marzorati (IEO); Monica Casiraghi (IEO); Gabriella Pravettoni (IEO)
Reviewers	iBeChange Consortium

Funded by
the European Union

This project has received funding from the European Union’s Horizon Europe research and innovation programme under the Grant Agreement Number 101136840.

Disclaimer

Table of Contents

List of Abbreviations

Abbreviation	Explanation
AUC	Area Under the Curve
BMI	Body Mass Index
CHAMP	Checklist for Statistical Assessment of Medical Papers
CI	Confidence Interval
COSMOS	Continuous Observation of SMOking Subjects
CRC	Colorectal Cancer
F1-score	Harmonic mean of precision and recall (performance metric)
FDR	False Discovery Rate
FIT	Fecal Immunochemical Test
FOBT	Fecal Occult Blood Test
GLMNet	Generalized Linear Model with Elastic Net Regularization
g/day	grams per day
HNPCC	Hereditary Non-Polyposis Colorectal Cancer
HRL	High-Risk Lesion
IBD	Inflammatory Bowel Disease
IQR	Interquartile Range
IRL	Intermediate-Risk Lesion
LDA	Linear Discriminant Analysis
LIME	Local Interpretable Model-agnostic Explanations
LRL	Low-Risk Lesion
MAR	Missing At Random
METs	Metabolic Equivalent of Task (hours per week)
MICE	Multivariate Imputation by Chained Equations
nnet	Neural Networks (algorithm)
OR	Odds Ratio
PRAUC	Precision-Recall Area Under the Curve
PSA	Prostate-Specific Antigen
ROC	Receiver Operating Characteristic
SD	Standard Deviation
SMOTE	Synthetic Minority Over-sampling Technique
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology
UI	User Interface
VIF	Variance Inflation Factor
WP3	Work Package 3
WP4	Work Package 4

XGBoost	Extreme Gradient Boosting (algorithm)

List of Tables

List of Figures

Executive Summary

This report presents the results of the work conducted under Task 3.1 of the *iBeChange* project, which aimed at gathering and exploiting retrospective data from colorectal and lung cancer screening programs to design, develop, and train models for future predictions and decision-making. The overarching goal was to understand the relationship between cancer onset and lifestyle and psychosocial risk factors, and to identify potential interventions to improve cancer outcomes.

A retrospective observational study was conducted using data from 1,074 participants in a Spanish colorectal cancer screening program. Additionally, data from 2,690 participants in an Italian lung cancer screening program were analyzed. A variety of demographic, behavioral, and psychosocial variables were evaluated using logistic regression and machine learning techniques (i.e., particularly Naïve Bayes models and Extreme Gradient Boosting) for their ability to predict colorectal and lung cancer risk.

The analysis of colorectal cancer data revealed marked sex-specific differences: in women, older age, diabetes, vocational education, and low dairy intake were key risk factors; in men, current and cumulative smoking, low nut intake, and higher alcohol consumption were most predictive. Importantly, psychological and motivational factors, such as readiness to reduce meat consumption, quit smoking, or increase physical activity, emerged as strong protective variables.

Lung cancer risk was strongly associated with smoking history (especially pack-years), respiratory symptoms like wheezing and coughing, and high consumption of processed meats. Key predictors included chronic bronchitis, wheezing, and smoking exposure, particularly with over 60 pack-years.

These findings, supported by different methodological approaches, highlight the value of incorporating behavioral readiness into intervention strategies.

1. Introduction

Work Package 3 (WP3) of the iBeChange project focuses on developing novel approaches for interaction through a Virtual User Model. The aim is to implement personalized and data-driven strategies that maximize user acceptance and adherence by collecting and analyzing retrospective and publicly available data related to behavioral and psychosocial risk factors. This information will inform the iBeChange Platform (WP4), contributing to the development of personalized interventions and interfaces to enhance user experience.

Deliverable 3.1 presents the results of the analysis of retrospective data under Task 3.1. The objective was to collect and analyze retrospective data from colorectal cancer (CRC) and lung cancer screening initiatives in order to design, build, and train predictive models to support future decision-making. The overarching goal of this task was to explore how lifestyle and psychosocial factors relate to the development of cancer and to identify possible strategies for improving cancer prevention and outcomes.

1.1 Colorectal Cancer Screening Programs

The global burden of cancer, in terms of both incidence and mortality, continues to rise, driven in part by population aging and growth, as well as shifts in the prevalence and distribution of key risk factors. CRC ranks as the third most commonly diagnosed malignancy and the second leading cause of cancer-related death worldwide, affecting both men and women.[1]

The overall 5-year survival rate for CRC ranges between 50% and 60%, with markedly higher rates observed in early stages (>90% in stage I; 60–85% in stage II) compared to advanced stages (25–65% in stage III; 5–7% in stage IV).[2]

While genetic predispositions—such as a family history of CRC or low-penetrance genetic variants—are established risk factors, the majority of CRC cases are attributable to modifiable lifestyle factors, including physical inactivity, alcohol consumption, tobacco use, and poor dietary habits.[3]

CRC is also preventable through population-based screening. Early detection offers the opportunity to identify the disease at a pre-malignant stage, before it progresses to an advanced and, consequently, less curable form.

Several methods can be employed for the early detection of CRC, but only two (fecal occult blood testing (FOBT) and sigmoidoscopy) have proved, in randomized controlled trials, to lower mortality rates.[4,5]

1.1.1 Colorectal Cancer Screening in Spain

In the present instance, the screening protocol in Spain is based on the FOBT. In fact, multiple trials indicate that FOBT screening cuts CRC deaths by roughly 15–25% (level of evidence 1a; recommendation grade A).[6] Moreover, by identifying and removing premalignant growths (most notably adenomatous polyps and serrated lesions), FOBT also helps reduce the overall incidence of CRC.

The primary goal of any CRC screening initiative is to catch cancer at an initial stage or to identify adenomas before they become malignant. Both early-stage CRC and these precancerous

adenomas tend to bleed intermittently at levels too small to be seen without testing, yet detectable by FOBT well before any clinical symptoms appear.

The target population consists of men and women between the ages of 50 and 69 residing anywhere within the territory of the Spanish state. There are two types of exclusions: definitive exclusion and temporary exclusion.

These are examples of some conditions that permanently exclude an individual from participating in the screening program: deceased, personal history of CRC, personal history of inflammatory bowel disease (IBD), including ulcerative colitis or Crohn’s disease, and colorectal adenomas, terminal illness or severe disability that contraindicates colon examination, personal history of total colectomy, family history of familial adenomatous polyposis or other polyposis syndromes, or hereditary non-polyposis CRC (HNPCC), family history of CRC, age error, or voluntary withdrawal.

The conditions that temporarily postpone participation in the screening program are colonoscopy performed within the past 5 years, and temporary illness or disability that does not contraindicate future testing.

The screening test used in the Spanish CRC screening programs is the quantitative immunochemical fecal occult blood test (FIT) for human hemoglobin, offered biennially.[7,8] This test involves collecting a small stool sample using a dedicated kit, which is then analyzed to detect the presence of occult (non-visible) blood.

Only one sample per participant is analyzed per screening round. The threshold for a positive FIT result is set at 20 µg of hemoglobin per gram of stool, equivalent to 100 ng/mL using the current analytical method.

If the FIT result is negative, the person will be reinvited after two years, assuming they remain eligible for screening. If the FIT result is positive, a colonoscopy is recommended to confirm or rule out the presence of a cancerous or premalignant lesion that may have caused the bleeding.

If the colonoscopy is normal or reveals only hyperplastic polyps, the screening process ends, and a new FIT is recommended in 10 years.

If the result shows a low-risk lesion (LRL), the individual will be reinvited for FIT screening in 2 years. In cases of intermediate-risk lesion (IRL), high-risk lesion (HRL), CRC, or other digestive diseases associated with increased CRC risk, the person is excluded from the CRC screening program and referred for follow-up through primary or specialized care as appropriate.

1.2 Lung Cancer Screening Programs

Lung cancer remains the leading cause of cancer-related mortality worldwide. According to recent global estimates, approximately 2.48 million new cases and 1.82 million deaths were attributed to lung cancer, corresponding to 12.4% of all new cancer diagnoses and 18.7% of cancer deaths globally.[9,10] While incidence rates are stabilizing or declining in many high-income countries due to reductions in smoking prevalence, they continue to rise in low- and middle-income regions, largely due to ongoing tobacco exposure and increasing air pollution (Tran et al., 2019). Early detection remains critical: when diagnosed at stage I, lung cancer has a 5-year survival rate exceeding 70%, compared to less than 10% for advanced-stage disease.[11]
These figures underscore the urgent need for organized screening strategies, yet many countries still lack a national lung cancer screening program. The most widely adopted screening methodology involves annual LDCT scans in individuals with a significant smoking history and other risk factors, sometimes supported by emerging tools such as biomarker-based risk stratification or individualized risk models to improve accuracy and cost-effectiveness.[12] Up to date, lung cancer screening programs, particularly those based on LDCT, have demonstrated efficacy in reducing lung cancer–specific mortality among high-risk populations.[11]

1.2.1 COSMOS: An Italian Study on the Continuous Observation of Smoking Subjects

In the context of secondary prevention of lung cancer, the European Institute of Oncology IRCCS (IEO) in Milan conducted two distinct prospective studies, COSMOS I and COSMOS II, aimed at evaluating and optimizing the effectiveness of low-dose computed tomography (LDCT) screening in high-risk individuals. The two studies have been here described separately, as they are characterized by different methodological designs and complementary specific objectives.

The **COSMOS I** (Continuous Observation of Smoking Subjects I) study was a landmark single-centre, prospective observational trial. Spanning from 2004 to 2015, COSMOS I aimed to evaluate the long-term efficacy of annual LDCT screening for the early detection of lung cancer in a high-risk, asymptomatic population.

A total of 5,207 individuals were enrolled, all meeting a defined high-risk profile: aged 50 years or older, with a minimum smoking history of 20 pack-years, and either current or former smokers. Key inclusion criteria also required that participants be asymptomatic for lung cancer, fit for potential surgical intervention, and with no history of cancer in the preceding five years, particularly no prior lung cancer. All participants provided written informed consent and agreed to undergo annual LDCT scans for a minimum of five years, with follow-up extending up to ten years. The study rigorously excluded individuals with severe comorbidities that would preclude curative treatment (e.g., end-stage COPD, advanced heart failure), those with prior lung cancer, a life expectancy of less than five years, or any contraindication to CT imaging, including severe claustrophobia or contrast allergies relevant to PET/CT. Additionally, participants needed to demonstrate the willingness and ability to comply with long-term follow-up protocols.

Over the course of the study, 1,035 volunteers were followed longitudinally for a full decade. During this time, 71 cases of lung cancer (6.9%) were diagnosed. Notably, the majority of these cancers (67%) were detected at stage I, underscoring the potential of LDCT screening to identify

lung cancer at a surgically curable phase. The survival outcomes were promising, with a five-year survival rate of 64% and a ten-year survival rate of 57% among those diagnosed.

The **COSMOS I study provided critical evidence supporting the feasibility and clinical utility of annual LDCT screening** in carefully selected high-risk populations. Its findings highlighted the value of early detection strategies in improving long-term survival in a disease often associated with late diagnosis and poor prognosis.

Following the promising results of the original COSMOS I trial, the **COSMOS II** study was initiated by the IEO as a multi-centre, prospective observational study. COSMOS II sought to expand, refine, and validate the early lung cancer detection strategies pioneered in COSMOS I, with a particular focus on enhancing screening precision and reducing the burden of unnecessary interventions.

A total of 3,107 participants were enrolled, representing a similar high-risk profile: asymptomatic individuals with significant smoking histories, eligible for curative treatment if needed. COSMOS II retained annual low-dose computed tomography (LDCT) as its primary screening modality but introduced an innovative biomarker-driven risk stratification approach. Central to the trial was the prospective validation of the "miR-Test", a serum microRNA signature previously developed during the COSMOS I study. This biomarker was used in combination with individualized risk models to personalize LDCT screening, identifying those most likely to benefit while reducing radiation exposure, healthcare costs, and overdiagnosis in lower-risk individuals.

Over ten years of follow-up, the study reported the detection of 297 lung cancers, corresponding to an incidence rate of approximately 7–8 cases per 1,000 person-years. Importantly, 76% of cancers were detected at stage I, reflecting a continued success in identifying early, potentially curable disease. Stage II, III, and IV cancers comprised 5%, 12%, and 7% of cases, respectively.

Outcomes were notably favorable: 89% of diagnosed patients underwent radical surgical resection, and the overall five-year survival rate approached 70%. Surgical procedures were generally safe, with perioperative mortality below 1%, highlighting the efficacy of the screening strategy in selecting operable candidates. **COSMOS II demonstrated the feasibility and clinical benefit of integrating molecular biomarkers with imaging in lung cancer screening**. By validating the miR-Test in a real-world, prospective setting and applying a personalized screening model, the study offered a roadmap for risk-adapted screening programs—balancing early detection with resource stewardship and patient safety. The findings from COSMOS II reinforce the role of precision medicine in population-based cancer screening and contribute to the evolution of lung cancer prevention strategies on a broader scale.

2. Methodology

This retrospective observational study aimed to develop predictive models for cancer outcomes, specifically CRC and lung cancer. We evaluated the association between the onset of these cancers and various lifestyle and psychosocial risk factors to identify potential interventions for improving outcomes. The study is reported in accordance with the STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guideline[13] and the CHAMP (Checklist for Statistical Assessment of Medical Papers) statement.[14] All analyses were performed using the R language.[15]

2.1 Data description and participants

To achieve this objective, two retrospective datasets were analyzed: one from a CRC screening program in Spain, and the other from the COSMOS lung cancer screening study in Italy.

Colorectal Cancer Screening Data. A subset of anonymized data was analyzed from participants in the CRC screening program in Spain, specifically from the northeastern region of the country, in the southern metropolitan area of Barcelona. This data was collected between 2016-2020 and includes information on 1074 participants. It contains detailed data on colonoscopy results and CRC diagnoses. The dataset also includes demographic information, including sex, age, and ethnic group, as well as various cancer risk factors, including family history of CRC, smoking, alcohol consumption, dietary habits, and sedentary lifestyle. A detailed description of the study participants can be found elsewhere.[16]

Lung Cancer Screening Data. A subset of anonymized data was analyzed from participants in the lung cancer screening program in Italy, collected between 2012 and 2016. The dataset includes information on 3,107 participants aged 50 or older with a heavy smoking history (≥ 20 pack-years) who were current or recent ex-smokers (quit within the past 10 years). Individuals with cancer diagnoses in the five years prior to enrollment were excluded. Data were collected using annual low-dose CT scans and self-reported questionnaires (T0 to T4), including information on medical history, respiratory symptoms, lifestyle, environmental exposures, and psychological status. Participants with a previous cancer diagnosis, or with a cancer diagnosis other than lung were excluded from the study.

2.2 Outcomes

Colorectal Cancer Screening. The primary outcome was the onset of CRC. Participants with a positive FIT result were referred for colonoscopy and subsequently classified into one of the following categories based on the findings: negative, polyps, LRL, IRL, HRL, or CRC. For the purpose of analysis, individuals diagnosed with HRL or CRC were grouped under the cancer outcome category, due to the strong association between HRLs and an increased risk of CRC development.[17] Conversely, the control group comprised participants with negative colonoscopy results or those diagnosed with polyps, LRL, or IRL.

Lung Cancer Screening. The primary outcome was defined as a new lung cancer diagnosis. Diagnoses were identified using a combination of self-reported responses collected at baseline and follow-up (T0 to T4) and data retrieved by data managers from clinical records. While participants were asked at each wave whether they had been diagnosed with cancer, this information was sometimes incomplete. Therefore, additional cases were included based on

clinical confirmation from annual low-dose CT (LDCT) scan records and diagnostic data curated by the COSMOS study data managers.

Participants who reported a cancer diagnosis at baseline (T0) were excluded from the outcome-positive group, except when the diagnosis referred to lung cancer identified through baseline LDCT screening. These cases were considered incidents due to the timing and nature of the screening process.

Importantly, only lung cancers were included as outcome events. Participants with other types of cancer (e.g., prostate, breast) were excluded to avoid misclassification and ensure consistency with the approach used in the colorectal cancer dataset. Likewise, participants with benign lung nodules (e.g., "Banning" cases) or collateral cancers not classified as lung cancer were also excluded. This strict definition aimed to focus the analysis on primary lung cancer cases and reduce misclassification bias.

The outcome variable was binary, indicating whether a lung cancer diagnosis was present (yes/no).

2.3 Predictors

The predictor variables encompassed lifestyle and psychosocial risk factors, as well as aspects of screening interventions and medical history.[18]

Lifestyle risk factors. Lifestyle risk factors for CRC were assessed through self-reported data on tobacco use, alcohol consumption, dietary habits, physical activity, and body weight. Additionally, participants completed a self-report measure evaluating their willingness to adopt lifestyle changes to reduce their risk of CRC. These questions explored participants' readiness to engage in behaviors such as adopting a healthier diet, exercising regularly, quitting smoking, reducing alcohol intake, and adjusting meat and vegetable consumption. By examining participants' intentions toward these modifications, the assessment provided valuable insights into their attitudes and motivation regarding CRC prevention.

Lifestyle predicting variables for lung cancer included age, sex, body mass index (BMI), tobacco use history (status and intensity), alcohol consumption, and dietary habits such as intake of fruits, vegetables, red and processed meat.

Psychosocial risk factors. Psychosocial risk factors for CRC included educational attainment, marital status, socioeconomic status, number of siblings and children, occupation, duration in the current role, type of employment, and self-reported history of depression or schizophrenia.

Perceived cancer risk was evaluated using self-reported measures that captured participants' cognitive and emotional responses to the possibility of developing cancer in the month preceding recruitment. Items assessed the frequency and impact of cancer-related thoughts on mood and daily functioning, the extent of worry regarding cancer development, and the personal significance attributed to these concerns. This assessment provided a nuanced understanding of participants' cancer risk perception and its psychological correlations.

Psychosocial predictors for lung cancer were measured using the Hospital Anxiety and Depression Scale (HADS), providing standardised assessments of anxiety and depressive symptoms.[19]

Screening interventions and medical history. Screening-related variables assessed in this study for CRC included participants' history of engagement in cancer screening programs within the past five years. Participants were also asked about their beliefs regarding the effectiveness of early detection programs in identifying breast or CRC at an early stage, as well as their intentions to participate in future CRC screening rounds.

Screening and medical history predictors for lung cancer included personal and family history of cancer, presence of comorbidities, medication use, participation in cancer screening exams, and exposure to occupational or environmental risk factors.

2.4 Statistical analysis

In the CRC dataset, all results were analyzed separately for men and women to account for sex-specific differences. This stratified approach was informed by marked disparities between sexes, as well as between cases and controls within each sex, allowing for a more precise understanding of these differences. To ensure consistency in the analysis and facilitate valid comparisons, identical statistical procedures were applied uniformly across both subgroups.

In contrast, the lung cancer dataset was not stratified by sex, as no significant sex-related differences were identified.

2.4.1 Exploratory analysis

For both datasets, the exploratory analysis began with a visual and descriptive assessment of all variables to evaluate frequencies, percentages, and near-zero variance for categorical variables (e.g., sex, education, occupation), as well as distributions and missing data patterns for continuous variables (e.g., age, weight).[20] Near-zero variance was defined as categorical variables with a low proportion of unique values relative to the sample size, indicating minimal variability; such cases were addressed by collapsing categories where appropriate.[21]

Missing data was handled using imputation methods, followed by sensitivity analyses to assess the robustness of findings with and without imputation. The Multivariate Imputation by Chained Equations (MICE) algorithm was employed, using the fully conditional specification method, in which each variable with missing data is imputed using a tailored model.[22,23] A total of 100 multiple imputations were performed using predictive mean matching for continuous variables, logistic regression for binary variables, and polytomous regression for categorical variables. Participants with missing outcome data across all follow-ups were excluded from the analysis.

In the lung cancer dataset, class imbalance was addressed through SMOTE (Synthetic Minority Over-sampling Technique), due to a significant difference in class sizes. We used the SMOTE function from the DMwR package in R, with a 50% oversampling rate, an undersampling rate of 500, and 5 nearest neighbors to generate new minority class examples.[24,25] This configuration increased the number of synthetic cancer cases while reducing the majority class to improve class balance without distorting the data distribution. This differs from the colorectal cancer dataset, where no SMOTE or resampling methods were applied, as class imbalance was not as pronounced. For the lung cancer dataset, all models were trained on the SMOTE-adjusted dataset to ensure proper representation of cancer cases. In contrast, the colorectal cancer dataset was modeled without any prior resampling, given the more balanced class distribution.

To examine inter-variable relationships, both datasets used correlation matrices and plots. Pearson, polychoric, and polyserial correlation coefficients were applied based on variable types (continuous, ordinal, or mixed). Descriptive statistics included means and standard deviations (SD) for normally distributed continuous variables, with group comparisons conducted using two-sample t-tests. For non-normally distributed variables, medians and interquartile ranges (IQR) were reported, and the Wilcoxon rank-sum test was used for comparisons. Normality was evaluated using the Shapiro–Wilk test. Categorical variables were summarized as frequencies and percentages, with group differences assessed via Chi-square tests. A two-sided p-value of < 0.05 was considered statistically significant.

2.4.2 Logistic regression models

The modeling approach consisted of two key steps. In the CRC dataset, the first step involved a series of logistic regression models to assess the association between CRC onset and individual lifestyle and psychosocial risk factors. Analyses were performed separately for men and women to account for sex-specific differences observed in the data. Each predictor was analyzed in an individual model, with all models adjusted for age. For diet-related predictors, total energy intake was additionally included as an adjustment to account for potential confounding.

In the lung cancer dataset, logistic regression models were also conducted to evaluate the association between cancer onset and relevant predictors. However, analyses were not stratified by sex, as no significant sex-related differences were identified in exploratory analyses. All models were adjusted for age.

In the second step, both datasets used multivariable logistic regression models that included a comprehensive set of predictors into a single model. To ensure model stability and minimize multicollinearity, variables with a pairwise correlation coefficient greater than 0.8 or a variance inflation factor (VIF) exceeding 5.0 were excluded.

Results were reported as odds ratios (OR) with 95% confidence intervals (CI). Associations were considered statistically significant if the CI did not include 1.0 or if the p-value was below 0.05. To control for the increased risk of false positives due to multiple comparisons, a Bonferroni correction was applied. The adjusted significance threshold was determined by dividing 0.05 by the number of tests performed, thereby reducing the likelihood of spurious findings.

2.4.3 Machine learning models

In both datasets, machine learning models were developed using lifestyle and psychosocial risk factors as features to predict cancer onset as the outcome. A 10-fold cross-validation approach was employed, with repeated sampling and replacement for model training and testing. To optimize model performance and reduce overfitting, nested resampling with subsampling was used to increase execution speed. This approach involved two stages: the outer resampling randomly selected 2/3 of the data for training (up to a maximum of 20,000 observations), with the remaining data used for validation. The inner resampling split the training data from the outer resampling into 2/3 for further training (up to a maximum of 10,000 observations) and the remaining data for testing.

The following classification models were used in both datasets: boosted trees, random forest, support vector machines, logistic regression, single-layer neural network, k-nearest neighbors, naive Bayes, and discriminant analysis. Model performance was compared using several

evaluation metrics, including area under the receiver operating characteristic (ROC) curve, sensitivity, specificity, Kappa, and positive and negative predictive values. The ROC curve plots sensitivity (y-axis) against 1-specificity (x-axis), with the area under the curve ranging from 0 to 1, where 1 indicates perfect prediction and 0.5 represents random chance.

2.4.4 Regression tree

In both datasets, regression trees (using recursive partitioning) were applied to the same set of predictors and outcomes. Regression trees offer a complementary approach by identifying optimal cut-points for predictor variables in relation to the outcome, while accounting for the influence of preceding splits. To mitigate the risk of overfitting, a cost-complexity pruning method was employed, using the weakest link pruning strategy. This technique involves iteratively collapsing the internal node that yields the smallest increase in the cost-complexity criterion.[26] Nodes were pruned only if overfitting was detected; otherwise, they were retained.

2.4.5 Interpretable machine learning models

In both datasets, to enhance the interpretability of individual predictions, a model-agnostic explanation technique, the Local Interpretable Model-Agnostic Explanations (LIME), was applied. This method identifies the contribution of individual risk factors to the prediction outcomes at the participant level.[27,28]

The LIME algorithm generates local explanations by introducing random noise to a participant's data to create a set of perturbed samples around the original observation. A weighted linear model is then fitted to this synthetic dataset, assigning greater weight to observations that are more similar to the one being explained.

2.5 Shiny application

Web-based applications were developed for both the CRC and lung cancer datasets using the open-source Shiny framework in the R programming language (https://shiny.rstudio.com/). The Shiny architecture facilitated the separation of the application into two main components: the user interface (UI) and the server.

The CRC-focused UI was designed as a questionnaire form, where users respond to multiple closed-ended questions corresponding to the predictor variables included in the CRC risk model. These predictors encompassed marital status, physical activity at work, METs hours per week, maximum weight, and age at maximum weight. Cancer-related psychological and behavioral variables included the impact of cancer-related thoughts on mood, as well as self-reported willingness to change lifestyle behaviors to CRC risk, such as losing weight if obese, increasing exercise if sedentary, and reducing meat intake if consuming a meat-heavy diet. Dietary intake variables included total protein, total carbohydrates, total ethanol consumption, and gram-per-day intakes of white meat, cured and processed meat, total meat, fruits, nuts, milk and yogurt, and caloric beverages. Smoking-related predictors included age at smoking initiation and smoking status. Clinical variables included laxative use and high cholesterol. All questions were implemented using single-choice radio buttons to ensure clarity and ease of response.

Upon completing the questionnaire, users can click the "Calculate your cancer risk factors" button, which sends their input to the server. On the server side, the user inputs are processed and passed through a pre-trained machine learning model for CRC risk prediction. The model

generates a LIME plot that highlights the most significant predictors of CRC risk for the specific user, based on their responses. The LIME plot is presented as a horizontal bar chart, visually breaking down the contributions of each predictor to the overall risk estimate. Positive (protective) and negative (risk-enhancing) factors are color-coded to facilitate immediate interpretation, helping users understand the key behaviors or characteristics influencing their predicted risk.

A similar Shiny application was developed for the lung cancer dataset, allowing users to quickly assess their risk factors through a user-friendly interface. Upon launching the app, users are guided through a series of categorical questions aligned with the variables included in the lung cancer predictive model. After completing the questionnaire and clicking the “Calculate your cancer risk factors” button, users receive a personalized bar chart showing the influence of each factor on their estimated cancer risk. An explanatory note is provided to aid interpretation, and a disclaimer emphasizes that the tool is predictive and not diagnostic. The app was designed for broad accessibility, with planned translations into Italian, Spanish, and Romanian to support multi-country deployment.

Although integration into the iBeChange platform was considered, the current predictive performance of the models led to further evaluation of the tool’s readiness for public health implementation.

3. Results

3.1 CRC screening program

3.1.1 Participants characteristics

Table 1 presents the characteristics of the 1074 participants in the CRC screening program cohort study, stratified by CRC diagnosis and gender. The cohort consists of 560 female participants, 430 controls and 76 cases; and 568 male participants, 420 with controls and 148 cases. The cohort was predominantly White/Caucasian (97.2% for females and 98.4% for males). Most participants were married or living with a partner, with 68.4% of females and 81.1% of males married or cohabiting in the CRC group.

Cases were significantly older than controls. Among females, the median age was 63.00 years [IQR: 57.75–65.25] for cases, compared to 60.00 years [IQR: 55.00–65.00] for controls (p = 0.009). Similarly, male cases had a higher median age at recruitment (63.00 years [IQR: 57.00–67.00]) than their control counterparts (60.00 years [IQR: 55.00–65.00], p = 0.005). This aligns with the well-established association between advancing age and increased cancer risk.[29]

Waist circumference was significantly associated with CRC among females, but not among males. Female cases had a significantly higher mean waist circumference compared to those without a CRC diagnosis (94.94 ± 11.40 cm vs. 91.47 ± 13.00 cm, p = 0.037).

Associations between cases and smoking-related variables were more pronounced among males. Male cases were significantly more likely to have ever smoked or to smoke regularly compared to those without CRC (87.8% vs. 76.2%, p = 0.004). A higher proportion of male cases also reported being current smokers (44.6 vs. 21.7%, p< 0.001) and to smoke daily (43.2% vs. 21.0%, p< 0.001). Additionally, male cases were less likely to be non-smokers (12.2% vs. 23.6%) or ex-smokers (43.2% vs. 54.7%, p< 0.001). The median number of pack-years was significantly higher among male cases (33.02 [IQR: 17.01–50.03] vs. male controls 25.82 [IQR: 11.41–45.03], p = 0.030). The median number of years of smoking was also significantly higher among male cases (38.50 [IQR: 30.00, 45.00]) than controls (31.00 [IQR: 20.00, 38.75], p< 0.001).

In terms of diabetes, a significantly higher proportion of female cases reported a history of diabetes (23.7% vs. 8.8%, p< 0.001), highlighting a potential link between diabetes and CRC risk in this cohort. However, the prevalence of diabetes was not significantly higher among male cases compared to controls.

Lifestyle modification intentions differed significantly between male cases and controls. Compared to controls, a smaller proportion of cases reported being willing to change their lifestyle to reduce CRC risk (91.9% vs. 97.8%, p = 0.021). However, a greater proportion of male cases expressed willingness to quit smoking if they were smokers (53.6% vs. 35.8%, p< 0.001) and to reduce alcohol consumption if they were heavy drinkers (62.0% vs. 45.8%, p= 0.005). In contrast, fewer male cases reported willingness to increase physical activity (88.0% vs. 94.2%, p= 0.026) or to reduce meat consumption (86.0% vs. 92.0%, p= 0.042).

Dietary patterns showed several significant associations with CRC, particularly among male participants. Among males, CRC cases had significantly higher legume consumption than controls, despite identical median intake values (38.57 g/day [IQR: 24.92–43.13] vs. 38.57 g/day [IQR: 26.67–51.43], p= 0.002), suggesting differences in the overall distribution. Male cases also

reported significantly lower nut intake (4.10 g/day [IQR: 0.98–15.00] vs. 6.43 g/day [IQR: 2.46–19.10], p= 0.024), a food group commonly associated with protective effects against cancer.[30,31] In contrast, among females, legume consumption did not differ significantly between CRC cases and controls. However, female cases consumed significantly less milk and yogurt compared to their control counterparts (205.17 g/day [IQR: 88.80–352.11] vs. 228.55 g/day [IQR: 160.90–369.00], p= 0.018). Additionally, alcohol consumption was notably higher among male cases (220.06 g/day [IQR: 65.33–354.14)] than controls (142.39 g/day [IQR: 34.11–306.71], p= 0.021), further highlighting potential lifestyle behavior differences between participants with and without CRC.

Supplemental Figure 1 displays the distribution of missing data across the dataset. In the chart, light blue areas indicate missing responses, particularly concentrated in the final two sections of the questionnaire: the food frequency questionnaire and the section assessing participants' attitudes toward cancer screening, cancer-related concerns, and willingness to adopt lifestyle changes to reduce CRC risk. Due to the high and systematic non-response in these sections, participants who left them incomplete were excluded from the analysis, resulting in the removal of 325 individuals from the cohort.

The remaining missing data were assumed to be missing at random (MAR) and were handled using multiple imputation techniques. Summary statistics comparing the imputed dataset to the original dataset are provided in **Supplemental Table 1**, demonstrating the consistency of key variables post-imputation.

Table 1. CRC screening program study sample characteristics.

	FEMALES				MALES			
Variables	**Controls (n=430)**	**Cases (n=76)**	**p-value**	**Missing (%)**	**Controls (n=420)**	**Cases (n=148)**	**p-value**	**Missing (%)**
Age at recruitment (median [IQR])	60 [5500, 65.00]	63.00 [57.75, 65.25]	0.009	0.0	60.00 [55.00, 65.00]	63.00 [57.00, 67.00]	0.005	0.0
Ethnicity: White/Caucasian (%)	418 (97.2)	74 (97.4)	1.000	0.0	411 (97.9)	148 (100.0)	0.158	0.0
Education level (%)			0.077	0.0			0.144	0.0
- University	54 (12.6)	2 (2.6)			73 (17.4)	28 (18.9)		
- High school diploma	68 (15.8)	13 (17.1)			69 (16.4)	31 (20.9)		
- Vocational training	82 (19.1)	19 (25.0)			100 (23.8)	26 (17.6)		
- Complete primary education	189 (44.0)	31 (40.8)			167 (39.8)	54 (36.5)		
- Incomplete primary education	30 (7.0)	8 (10.5)			10 (2.4)	7 (4.7)		
- No formal education, but can read	7 (1.6)	3 (3.9)			1 (0.2)	2 (1.4)		
Marital status (%)			0.180	0.2			0.387	0.0
- Single/never married	27 (6.3)	4 (5.3)			26 (6.2)	5 (3.4)		
- Married or living with a partner	329 (76.7)	52 (68.4)			345 (82.1)	120 (81.1)		
- Separated or divorced	45 (10.5)	10 (13.2)			42 (10.0)	19 (12.8)		
- Widowed	28 (6.5)	10 (13.2)			7 (1.7)	4 (2.7)		
Social class of parents (%)			0.425	0.6			0.996	0.2
- Upper social class	5 (1.2)	0 (0.0)			6 (1.4)	2 (1.4)		
- Middle social class	240 (56.2)	39 (51.3)			228 (54.4)	81 (54.7)		
- Lower social class	182 (42.6)	37 (48.7)			185 (44.2)	65 (43.9)		

Number of siblings (median [IQR])	3.00 [1.00, 4.75]	2.00 [1.00, 4.00]	0.414	0.2	2.00 [1.00, 4.00]	2.00 [1.00, 4.00]	0.239	0.0
Number of children (median [IQR])	2.00 [1.00, 2.00]	2.00 [1.00, 2.00]	0.135	0.2	2.00 [1.00, 2.00]	2.00 [1.00, 2.00]	0.601	0.0
Weight (median [IQR])	68.00 [60.00, 77.00]	70.00 [62.00, 77.25]	0.348	0.2	83.00 [75.00, 92.25]	82.50 [74.00, 93.00]	0.485	0.0
Occupation (%)			0.269	0.0			0.472	0.2
- Working	179 (41.6)	26 (34.2)			203 (48.3)	63 (42.9)		
- Unemployed	61 (14.2)	10 (13.2)			27 (6.4)	12 (8.2)		
- Housewife or domestic worker	78 (18.1)	21 (27.6)						
- Retired	112 (26.0)	19 (25.0)			190 (45.2)	72 (49.0)		
Physical activity at work (%)			0.192	2.8			0.092	0.2
- Sedentary	65 (15.4)	11 (15.7)			61 (14.5)	17 (11.6)		
- Slightly active	80 (19.0)	6 (8.6)			65 (15.5)	18 (12.2)		
- Moderately active	93 (22.0)	14 (20.0)			109 (26.0)	50 (34.0)		
- Fairly active	129 (30.6)	29 (41.4)			130 (31.0)	35 (23.8)		
- Very active	55 (13.0)	10 (14.3)			55 (13.1)	27 (18.4)		
METs hours per week (median [IQR])	14.25 [0.00, 24.00]	20.80 [0.00, 38.72]	0.126	0.0	21.00 [6.00, 40.80]	15.85 [0.00, 36.60]	0.084	1.2
METs hours per week walking (median [IQR])	0.00 [0.00, 15.00]	0.00 [0.00, 21.00]	0.407	0.0	0.00 [0.00, 21.00]	0.00 [0.00, 21.00]	0.557	0.2
Waist circumference (mean (SD))	91.47 (13.00)	94.94 (11.40)	0.037	9.5	99.29 (10.92)	100.43 (12.31)	0.323	10.6
Hip circumference (median [IQR])	103.00 [96.00, 109.00]	105.00 [99.50, 112.00]	0.083	12.1	103.00 [98.00, 108.00]	102.00 [98.00, 108.00]	0.566	15.3
Waist-hip ratio (median [IQR])	0.89 [0.83, 0.93]	0.89 [0.84, 0.95]	0.140	12.1	0.96 [0.92, 1.02]	0.98 [0.94, 1.03]	0.088	15.3
Weight 1 year ago (median [IQR])	67.00 [60.00, 76.00]	70.00 [60.00, 78.50]	0.374	1.8	83.00 [75.00, 93.25]	82.00 [73.75, 93.00]	0.583	0.7

Maximum weight (median [IQR])	71.00 [64.00, 82.00]	74.00 [65.50, 83.50]	0.261	1.0	88.00 [79.00, 97.00]	86.50 [80.00, 98.00]	0.579	0.7
Age at maximum weight (median [IQR])	56.00 [50.00, 63.00]	57.00 [50.00, 64.00]	0.306	1.4	55.00 [50.00, 62.00]	59.00 [50.00, 64.75]	0.092	0.5
Current height (cm) (median [IQR])	157.00 [153.00, 162.00]	158.00 [153.75, 160.00]	0.992	0.2	171.00 [166.00, 175.00]	170.00 [167.00, 175.00]	0.402	0.0
BMI (median [IQR])	27.24 [24.44, 30.85]	28.25 [25.54, 30.95]	0.288	0.4	28.37 [26.09, 31.60]	27.89 [25.93, 30.86]	0.527	0.0
In your lifetime, have you ever smoked? 'YES' means at least 100 cigarettes or 360 grams of tobacco in your lifetime. (%)	204 (47.4)	38 (50.0)	0.774	0.0	320 (76.2)	130 (87.8)	0.004	0.0
Have you ever smoked regularly, i.e., at least one cigarette per day for six months or more? (%)	203 (47.2)	38 (50.0)	0.746	0.0	320 (76.2)	130 (87.8)	0.004	0.0
Age at smoking initiation (median [IQR])	17.00 [15.00, 19.00]	18.00 [15.25, 23.00]	0.082	0	17.00 [15.00, 18.25]	17.00 [15.00, 18.00]	0.691	0.2
Current smoker (%)	87 (20.3)	16 (21.1)	1.000	0.2	91 (21.7)	66 (44.6)	<0.001	0.0
Number of cigarettes on average, excluding non-smokers (median [IQR])	12.00 [9.00, 20.00]	15.00 [13.75, 20.00]	0.315	27.3	15.00 [7.50, 20.00]	10.00 [5.00, 20.00]	0.171	51.6
Current frequency of smoking (%)			0.913	0.2			<0.001	0.0
- Day	83 (19.3)	16 (21.1)			88 (21.0)	64 (43.2)		
- Week	3 (0.7)	0 (0.0)			2 (0.5)	2 (1.4)		
- Month	1 (0.2)	0 (0.0)			1 (0.2)	0 (0.0)		
- Former smoker	116 (27.0)	22 (28.9)			229 (54.5)	64 (43.2)		
- Never	226 (52.7)	38 (50.0)			100 (23.8)	18 (12.2)		

Passive smoker (%)	114 (31.1)	21 (32.8)	0.905	15.0	88 (26.4)	37 (32.5)	0.264	21.3
Smoking status (%)			0.901	0.0			<0.001	0.2
- Never	227 (52.8)	38 (50.0)			99 (23.6)	18 (12.2)		
- Ex-Smoker	116 (27.0)	22 (28.9)			229 (54.7)	64 (43.2)		
- Smoker	87 (20.2)	16 (21.1)			91 (21.7)	66 (44.6)		
Pack years, excluding never-smokers (median [IQR])	19.31 [5.03, 34.02]	25.52 [15.71, 39.03]	0.212	1.2	25.82 [11.41, 45.03]	33.02 [17.01, 50.03]	0.030	5.6
Average lifetime intensity in cigarettes/year, excluding never-smokers (median [IQR])	4383.00 [2192.00, 7305.00]	5478.00 [3652.00, 7305.00]	0.450	0.2	7305.00 [3652.00, 10958.00]	7305.00 [3652.00, 10945.50]	0.381	4.9
Years of smoking, excluding non-smokers (median [IQR])	33.00 [20.00, 39.00]	34.00 [23.00, 39.00]	0.643	1.2	31.00 [20.00, 38.75]	38.50 [30.00, 45.00]	<0.001	2.1
Heartburn (%)	182 (42.4)	27 (35.5)	0.318	0.2	136 (32.5)	42 (28.8)	0.469	0.5
Medication for heartburn (%)	179 (41.6)	27 (35.5)	0.384	0.0	155 (36.9)	45 (30.4)	0.186	0.0
Laxative use (%)	121 (28.5)	21 (28.4)	1.000	1.6	44 (10.6)	8 (5.5)	0.098	0.9
Diabetes = Yes (%)	38 (8.8)	18 (23.7)	<0.001	0.0	63 (15.0)	20 (13.5)	0.760	0.0
Hypertension = Yes (%)	136 (31.6)	26 (34.2)	0.755	0.0	180 (42.9)	72 (48.6)	0.261	0.0
High cholesterol = Yes (%)	136 (31.6)	32 (42.1)	0.098	0.0	167 (39.9)	65 (43.9)	0.443	0.2
Angina pectoris (%)	7 (1.6)	0 (0.0)	0.557	0.0	14 (3.3)	5 (3.4)	1.000	0.0
Myocardial infarction (%)	5 (1.2)	0 (0.0)	0.752	0.0	14 (3.3)	9 (6.1)	0.224	0.0
Stroke (%)	11 (2.6)	2 (2.6)	1.000	0.0	16 (3.8)	6 (4.1)	1.000	0.0
Circulatory problems (%)	53 (12.3)	6 (7.9)	0.360	0.0	38 (9.1)	14 (9.5)	1.000	0.4
Arthritis (%)	128 (29.8)	28 (36.8)	0.279	0.2	79 (18.9)	29 (19.6)	0.940	0.2
Migraine (%)	68 (15.8)	10 (13.2)	0.675	0.0	23 (5.5)	9 (6.1)	0.951	0.2
Anemia (%)	55 (12.8)	8 (10.5)	0.717	0.0	11 (2.6)	3 (2.0)	0.933	0.4

Diverticulitis (%)	5 (1.2)	2 (2.7)	0.622	0.2	8 (1.9)	2 (1.4)	0.944	0.4
Celiac disease (%)	5 (1.2)	0 (0.0)	0.750	0.4	0 (0.0)	0 (0.0)		
Depression (%)	126 (29.4)	26 (34.7)	0.439	0.6	53 (12.6)	23 (15.6)	0.438	0.4
Osteoporosis (%)	51 (11.9)	11 (14.5)	0.652	0.0	1 (0.2)	0 (0.0)	1.000	0.2
Polyps (%)	17 (4.0)	4 (5.3)	0.829	0.0	31 (7.4)	6 (4.1)	0.224	0.0
Dyspepsia (%)	27 (6.3)	6 (7.9)	0.792	0.4	15 (3.6)	1 (0.7)	0.123	0.0
Schizophrenia (%)	1 (0.2)	0 (0.0)	1.000	0.6	2 (0.5)	0 (0.0)	0.973	0.0
Anti-inflammatory medication (%)	111 (27.4)	21 (29.2)	0.869	5.7	81 (20.7)	30 (21.3)	0.974	6.2
Menstruation status = still has periods (%)	26 (6.1)	1 (1.3)	0.161	0.6				
Age at last menstruation (median [IQR])	50.00 [47.00, 52.00]	50.00 [45.00, 52.00]	0.387	10.3				
Age at first menstruation (%)			0.114	1.6				
- 8 - 11	111 (26.2)	27 (36.0)						
- 11 - 13	172 (40.7)	33 (44.0)						
- 13 - 14	99 (23.4)	10 (13.3)						
- 14 - 18	41 (9.7)	5 (6.7)						
Use contraceptive (%)	294 (69.2)	47 (62.7)	0.326	1.2				
Menopause treatment (%)	50 (11.9)	11 (15.5)	0.514	3.0				
Prostate disease (%)					81 (19.4)	26 (17.7)	0.743	0.5
Weight loss (%)	6 (1.4)	1 (1.3)	1.000	0.0	4 (1.0)	0 (0.0)	0.535	0.0
Are early detection programs useful? (%)			0.181	35.0			0.501	32.7
- Strongly disagree	2 (0.7)	0 (0.0)			5 (1.8)	2 (2.0)		
- Agree	21 (7.3)	0 (0.0)			18 (6.4)	6 (5.9)		
- Strongly agree	266 (92.0)	40 (100.0)			255 (90.7)	90 (89.1)		

Willing to participate again? (%)	277 (99.6)	35 (97.2)	0.547	37.9	269 (99.3)	95 (99.0)	1.000	35.4
During the past month, how often have you thought about your chances of getting cancer? (%)			0.717	34.6			0.222	32.2
- Rarely or never	98 (33.9)	11 (26.2)			127 (44.9)	34 (33.3)		
- Sometimes	138 (47.8)	23 (54.8)			132 (46.6)	56 (54.9)		
- Often	42 (14.5)	7 (16.7)			19 (6.7)	9 (8.8)		
- Disagree					0 (0.0)	1 (1.0)		
- Neither agree nor disagree					2 (0.7)	2 (2.0)		
- Almost all the time	11 (3.8)	1 (2.4)			5 (1.8)	3 (2.9)		
During the past month, has thinking about the possibility of developing cancer affected your mood? (%)			0.701	34.6			0.400	32.4
- NS/NC (Not sure/No comment)					1 (0.4)	0 (0.0)		
- Rarely or never	155 (53.6)	21 (50.0)			178 (62.9)	56 (55.4)		
- Sometimes	104 (36.0)	17 (40.5)			91 (32.2)	41 (40.6)		
- Often	23 (8.0)	4 (9.5)			12 (4.2)	4 (4.0)		
- Almost all the time	7 (2.4)	0 (0.0)			2 (0.7)	0 (0.0)		
During the past month, has thinking about the possibility of developing cancer affected your ability to carry out your daily activities? (%)			0.785	34.6			0.706	32.6
- Rarely or never	173 (59.9)	26 (61.9)			198 (70.2)	67 (66.3)		

- Sometimes	96 (33.2)	13 (31.0)			75 (26.6)	31 (30.7)		
- Often	15 (5.2)	3 (7.1)			7 (2.5)	3 (3.0)		
- Almost all the time	5 (1.7)	0 (0.0)			2 (0.7)	0 (0.0)		
To what extent do you worry about the possibility of developing cancer one day? (%)			0.449	34.6			0.587	32.4
- Not at all	70 (24.2)	9 (21.4)			82 (29.1)	35 (34.3)		
- A little	118 (40.8)	13 (31.0)			130 (46.1)	40 (39.2)		
- Quite a bit	72 (24.9)	14 (33.3)			51 (18.1)	18 (17.6)		
- A great deal	29 (10.0)	6 (14.3)			19 (6.7)	9 (8.8)		
How often do you worry about the possibility of developing cancer? (%)			0.809	34.6			0.550	32.4
- Never or rarely	102 (35.3)	17 (40.5)			120 (42.6)	44 (43.1)		
- Occasionally	156 (54.0)	20 (47.6)			147 (52.1)	49 (48.0)		
- Frequently	29 (10.0)	5 (11.9)			13 (4.6)	7 (6.9)		
- Constantly	2 (0.7)	0 (0.0)			2 (0.7)	2 (2.0)		
Is being worried about developing cancer an important issue for you? (%)			0.989	34.6			0.895	32.4
- No; not at all	92 (31.8)	13 (31.0)			115 (40.8)	46 (45.1)		
- A little	81 (28.0)	12 (28.6)			65 (23.0)	21 (20.6)		
- Yes; it's definitely a problem	73 (25.3)	10 (23.8)			67 (23.8)	23 (22.5)		
- Yes; it's a very serious problem	43 (14.9)	7 (16.7)			35 (12.4)	12 (11.8)		
Willing to change the lifestyle to reduce colon cancer risk (%)	271 (97.1)	40 (100.0)	0.586	37.0	264 (97.8)	91 (91.9)	0.021	35.0

If you were obese, would you lose weight? (%)			0.733	35.2			0.239	34.2
- Yes	216 (75.0)	31 (77.5)			212 (77.4)	72 (72.0)		
- No	5 (1.7)	0 (0.0)			3 (1.1)	4 (4.0)		
- I'm not obese	63 (21.9)	9 (22.5)			58 (21.2)	24 (24.0)		
- Not sure	4 (1.4)	0 (0.0)			1 (0.4)	0 (0.0)		
If you were a smoker, would you quit smoking? (%)			0.508	35.4			<0.001	35.2
- Yes	86 (30.0)	16 (40.0)			97 (35.8)	52 (53.6)		
- No	12 (4.2)	1 (2.5)			6 (2.2)	9 (9.3)		
- I'm not a smoker	184 (64.1)	23 (57.5)			165 (60.9)	34 (35.1)		
- Not sure	5 (1.7)	0 (0.0)			3 (1.1)	2 (2.1)		
If you were a heavy drinker, would you reduce your alcohol consumption? (%)			0.759	35.2			0.005	34.3
- Yes	81 (28.1)	11 (27.5)			125 (45.8)	62 (62.0)		
- No	2 (0.7)	0 (0.0)			5 (1.8)	5 (5.0)		
- I drink less alcohol	199 (69.1)	29 (72.5)			138 (50.5)	31 (31.0)		
- Not sure	6 (2.1)	0 (0.0)			5 (1.8)	2 (2.0)		
If you did little exercise: would you do more exercise on a regular basis? (%)			0.535	35.2			0.026	34.2
- Yes	259 (89.9)	38 (95.0)			258 (94.2)	88 (88.0)		
- No	7 (2.4)	0 (0.0)			6 (2.2)	8 (8.0)		
- I exercise a lot	14 (4.9)	2 (5.0)			6 (2.2)	4 (4.0)		
- Not sure	8 (2.8)	0 (0.0)			4 (1.5)	0 (0.0)		
If you were to eat a meat-heavy diet: would you eat less meat? (%)			0.407	35.2			0.042	34.2

- Yes	237 (82.3)	30 (75.0)			252 (92.0)	86 (86.0)		
- No	2 (0.7)	0 (0.0)			2 (0.7)	5 (5.0)		
- I don't eat much meat	45 (15.6)	10 (25.0)			19 (6.9)	9 (9.0)		
- Not sure	4 (1.4)	0 (0.0)			1 (0.4)	0 (0.0)		
If you were to eat a diet low in vegetables: would you eat more vegetables? (%)			0.576	35.2			0.292	34.2
- Yes	243 (84.4)	32 (80.0)			247 (90.1)	91 (91.0)		
- No	4 (1.4)	1 (2.5)			6 (2.2)	5 (5.0)		
- I eat a lot of vegetables	35 (12.2)	7 (17.5)			19 (6.9)	4 (4.0)		
- Not sure	6 (2.1)	0 (0.0)			2 (0.7)	0 (0.0)		
Total energy (kcal/day) (median [IQR])	1576.80 [1277.66, 1940.25]	1473.67 [1316.56, 1692.40]	0.266	8.7	1906.89 [1499.86, 2361.09]	1952.05 [1560.55, 2331.48]	0.912	8.1
Total protein (g/day) (median [IQR])	70.73 [56.19, 84.43]	68.97 [60.86, 80.61]	0.610	8.7	84.16 [67.73, 101.33]	83.22 [70.59, 98.80]	0.810	8.1
Total carbohydrates (g/day) (median [IQR])	150.87 [117.97, 192.87]	143.69 [125.45, 171.00]	0.458	8.7	184.56 [143.75, 229.36]	176.46 [136.40, 229.30]	0.423	8.1
Total fats (g/day) (median [IQR])	69.58 [52.06, 90.49]	67.49 [52.25, 79.46]	0.320	8.7	76.43 [59.29, 99.45]	80.84 [58.41, 96.76]	0.963	8.1
Total fiber (g/day) (median [IQR])	17.86 [13.89, 24.04]	18.47 [14.50, 24.59]	0.442	8.7	17.85 [13.95, 23.34]	17.47 [13.93, 22.03]	0.250	8.1
Total ethanol (g/day) (median [IQR])	1.53 [0.00, 6.70]	2.20 [0.00, 7.90]	0.608	8.7	10.30 [2.85, 22.51]	14.60 [4.92, 30.32]	0.007	8.1
Red meat (g/day) (median [IQR])	16.78 [7.13, 27.65]	15.35 [7.69, 29.16]	0.743	8.7	29.03 [17.53, 45.64]	29.56 [16.53, 42.83]	0.677	8.1
White meat (g/day) (median [IQR])	18.76 [13.31, 35.80]	19.02 [13.45, 33.72]	0.797	8.7	23.45 [18.16, 42.09]	24.05 [17.59, 40.28]	0.881	8.1
Cured and processed meat (g/day) (median [IQR])	24.80 [14.35, 38.06]	27.57 [15.28, 38.69]	0.937	8.7	42.15 [27.19, 61.62]	42.11 [28.87, 63.54]	0.617	8.1

All meat (g/day) (median [IQR])	72.36 [48.81, 96.01]	72.74 [47.81, 91.17]	0.746	8.7	105.40 [78.22, 141.40]	101.80 [78.31, 141.26]	0.984	8.1
White fish (g/day) (median [IQR])	14.44 [6.04, 18.74]	15.76 [9.63, 23.39]	0.073	8.7	15.73 [6.39, 19.11]	15.35 [6.29, 20.88]	0.981	8.1
Blue fish (g/day) (median [IQR])	13.63 [3.29, 18.08]	15.35 [6.33, 20.44]	0.134	8.7	15.43 [6.04, 19.84]	15.42 [3.37, 20.66]	0.877	8.1
Fruits (g/day) (median [IQR])	225.35 [133.33, 359.24]	253.78 [151.70, 367.41]	0.488	9.7	183.79 [100.21, 303.53]	168.55 [85.57, 258.18]	0.177	8.5
Vegetables (g/day) (median [IQR])	169.14 [119.16, 245.22]	192.45 [135.93, 266.87]	0.061	9.1	130.45 [87.66, 198.28]	127.51 [82.83, 203.30]	0.544	8.1
Legumes (g/day) (median [IQR])	36.65 [21.13, 40.97]	36.65 [25.33, 40.97]	0.838	9.1	38.57 [24.92, 43.13]	38.57 [26.67, 51.43]	0.002	8.1
Nuts (g/day) (median [IQR])	6.43 [0.98, 17.14]	6.43 [2.46, 17.14]	0.635	8.7	6.43 [2.46, 19.10]	4.10 [0.98, 15.00]	0.024	8.1
Dairy and desserts (g/day) (median [IQR])	7.02 [2.08, 20.61]	6.43 [2.83, 26.97]	0.351	8.7	11.91 [3.28, 31.36]	12.00 [3.71, 30.78]	0.781	8.1
Cheese (g/day) (median [IQR])	17.36 [6.73, 34.07]	17.58 [7.28, 32.81]	0.882	8.7	15.00 [6.43, 27.91]	14.17 [6.43, 26.53]	0.332	8.1
Milk and yogurt (g/day) (median [IQR])	228.55 [160.90, 369.00]	205.17 [88.80, 352.11]	0.018	9.7	226.34 [137.04, 388.05]	225.00 [101.83, 356.29]	0.114	9.3
Caloric beverages (g/day) (median [IQR])	19.68 [0.00, 66.62]	19.68 [0.00, 150.00]	0.851	9.1	53.50 [5.90, 164.20]	39.44 [0.00, 186.12]	0.827	9.0
Alcoholic beverages (g/day) (median [IQR])	20.34 [0.00, 99.08]	29.23 [0.00, 121.40]	0.622	8.7	142.39 [34.11, 306.71]	220.06 [65.33, 354.14]	0.021	9.5

CRC: Colorectal Cancer; IQR: Interquartile Range; METs: Metabolic Equivalent of Task; BMI: Body Mass Index. Two-sample t-tests were used for normally distributed continuous variables, Wilcoxon rank-sum tests were used for non-normally distributed continuous variables, and Chi-square tests were used for categorical variables.

3.1.2 Logistic Regression Models

Table 2 and **Table 3** summarize the results of the logistic regression models evaluating individual predictors of CRC risk, stratified by sex. **Table 2** presents results for women, and **Table 3** for men. Each predictor was assessed in a separate model, with all models adjusted for age. For models

involving dietary predictors, additional adjustment for total energy intake was applied to account for potential confounding. Results are reported as odds ratios (ORs) with corresponding 95% confidence intervals (CIs) and p-values.

3.1.2.1 Simple logistic regression models: Females

Table 2 presents findings from simple logistic regression models assessing CRC risk predictors among female participants. Among women, physical activity levels (as measured in MET-hours per week) were initially associated with CRC incidence. Compared to inactive individuals (0 METs/week), those engaging in low levels of physical activity (0.01–17.4 METs/week) demonstrated significantly reduced odds of developing CRC (OR = 0.437; 95% CI: 0.197–0.926; p= 0.035). However, this association did not remain statistically significant after adjustment for multiple comparisons using the false discovery rate (FDR) correction (adjusted p= 0.472). No other physical activity categories showed significant associations with CRC risk.

A similar pattern was observed for diabetes status. Women with diabetes exhibited higher odds of CRC compared to non-diabetic participants (OR = 2.49; 95% CI: 1.26–4.83; p= 0.007), but this association also lost statistical significance after FDR correction (adjusted p= 0.242).

Smoking-related variables showed consistent associations with increased CRC risk. When cumulative tobacco exposure was measured in cigarettes per year, women with moderate exposure (3,653–7,305 cigarettes/year) had significantly higher odds of CRC compared to never-smokers (OR = 2.44; 95% CI: 1.19–4.97; p= 0.014); however, this finding was not statistically significant after multiple testing adjustment (adjusted p= 0.31). Similar results were observed when smoking exposure was assessed using pack-years: participants with moderate-to-high exposure (25.23–42.03 pack-years) had increased odds of CRC (OR = 2.57; 95% CI: 1.16–5.62; p= 0.018), though the association did not persist after correction (adjusted p= 0.31).

Regarding dietary factors, milk and yogurt consumption appeared to confer a protective effect. Women in the highest intake category (369.01–885.42 g/day) had significantly lower odds of CRC compared to those in the lowest intake group (0–125 g/day) (OR = 0.224; 95% CI: 0.083–0.541; p= 0.002). Nonetheless, this association also failed to retain significance after FDR correction (adjusted p= 0.138). No other dietary variables were significantly associated with CRC onset.

Table 2. Simple logistic regression models results for the female subgroup.

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
METs hours per week				
- 0	1 [Reference]			Cases: 24 Controls: 72
- 0.01 - 17.4	0.437 (0.197, 0.926)	0.035	0.472	Cases: 12 Controls: 82
- 17.41 - 31	0.705 (0.342, 1.43)	0.335	0.725	Cases: 18 Controls: 67
- 31.01 - 140	1.38 (0.682, 2.8)	0.365	0.725	Cases: 22 Controls: 43
METs hours per week walking				
- 0	1 [Reference]			Cases: 42 Controls: 139
- 0.01 - 18	0.49 (0.22, 1)	0.063	0.472	Cases: 10 Controls: 70
- 18.01 - 108	1.3 (0.704, 2.36)	0.394	0.725	Cases: 24 Controls: 55
Waist circumference				
- 0 - 88	1 [Reference]			Cases: 29 Controls: 119
- 88.01 - 96	0.844 (0.411, 1.68)	0.634	0.795	Cases: 15 Controls: 68
- 96.01 - 104	1.67 (0.818, 3.34)	0.154	0.559	Cases: 19 Controls: 39
- 104.01 - 137	1.18 (0.533, 2.5)	0.677	0.834	Cases: 13 Controls: 38
Age at smoking initiation				
- 8 - 15	1 [Reference]			Cases: 10 Controls: 34
- 15 - 17	0.727 (0.222, 2.22)	0.582	0.787	Cases: 6 Controls: 30
- 17 - 19	0.612 (0.184, 1.89)	0.402	0.725	Cases: 6 Controls: 29
- 19 - 54	1.52 (0.592, 4.04)	0.387	0.725	Cases: 16 Controls: 30
- Never smoked	0.636 (0.279, 1.52)	0.29	0.725	Cases: 38 Controls: 141
Diabetes				
- No	1 [Reference]			Cases: 58 Controls: 236

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
- Yes	2.49 (1.26, 4.83)	0.007	0.242	Cases: 18 Controls: 28
Anti-inflammatory medication				
- No	1 [Reference]			Cases: 55 Controls: 199
- Yes	1.21 (0.665, 2.15)	0.524	0.749	Cases: 21 Controls: 65
In your lifetime, have you ever smoked? 'YES' means at least 100 cigarettes or 360 grams of tobacco in your lifetime.				
- No	1 [Reference]			Cases: 38 Controls: 140
- Yes	1.53 (0.881, 2.67)	0.133	0.54	Cases: 38 Controls: 124
Have you ever smoked regularly, i.e., at least one cigarette per day for six months or more?				
- No	1 [Reference]			Cases: 38 Controls: 141
- Yes	1.57 (0.902, 2.75)	0.113	0.517	Cases: 38 Controls: 123
Current smoker				
- No	1 [Reference]			Cases: 60 Controls: 219
- Yes	1.65 (0.83, 3.18)	0.143	0.548	Cases: 16 Controls: 45
Current frequency of smoking				
- Day or week	1 [Reference]			Cases: 16 Controls: 45
- Former smoker	0.738 (0.347, 1.59)	0.431	0.725	Cases: 22 Controls: 78
- Never	0.528 (0.257, 1.1)	0.083	0.472	Cases: 38 Controls: 141
Smoking status				
- Never	1 [Reference]			Cases: 38 Controls: 141
- Ex-Smoker	1.4 (0.74, 2.62)	0.296	0.725	Cases: 22 Controls: 78
- Smoker	1.9 (0.908, 3.89)	0.083	0.472	Cases: 16 Controls: 45

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
Passive smoker				
- No	1 [Reference]			Cases: 52 Controls: 193
- Yes	1.18 (0.664, 2.06)	0.565	0.78	Cases: 24 Controls: 71
Average annual cigarettes during the time smoked				
- Never smoked	1 [Reference]			Cases: 38 Controls: 141
- 0 - 3652	1.46 (0.721, 2.91)	0.285	0.725	Cases: 17 Controls: 62
- 3652 - 7305	1.35 (0.525, 3.18)	0.512	0.749	Cases: 8 Controls: 27
- 7305 - 29220	1.96 (0.886, 4.21)	0.089	0.472	Cases: 13 Controls: 34
Average lifetime intensity in cigarettes/year				
- Never smoked	1 [Reference]			Cases: 38 Controls: 141
- 36 - 3652	1.14 (0.557, 2.28)	0.709	0.843	Cases: 15 Controls: 66
- 3653 - 7305	2.44 (1.19, 4.97)	0.014	0.31	Cases: 18 Controls: 38
- 7306 - 9131	1.28 (0.063, 9.28)	0.832	0.934	Cases: 1 Controls: 4
- 9132 - 29220	1.38 (0.367, 4.26)	0.597	0.787	Cases: 4 Controls: 15
Years of smoking				
- Never smoked	1 [Reference]			Cases: 38 Controls: 141
- 1–21 years	1.36 (0.525, 3.28)	0.504	0.749	Cases: 8 Controls: 34
- 22–40 years	1.86 (0.96, 3.61)	0.064	0.472	Cases: 22 Controls: 64
- 41+ years	1.25 (0.491, 2.95)	0.615	0.787	Cases: 8 Controls: 25
Pack years				
- Never smoked	1 [Reference]			Cases: 38 Controls: 141
- 0.09 - 11.01	1.03 (0.426, 2.31)	0.947	0.978	Cases: 9 Controls: 46

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
- 11.02 - 25.22	1.62 (0.704, 3.59)	0.239	0.725	Cases: 11 Controls: 35
- 25.23 - 42.03	2.57 (1.16, 5.62)	0.018	0.31	Cases: 14 Controls: 28
- 42.04 - 168.12	1.23 (0.33, 3.74)	0.731	0.855	Cases: 4 Controls: 14
BMI				
- Normal or underweight (< 25)	1 [Reference]			Cases: 18 Controls: 84
- Overweight (25 - 29.9)	1.34 (0.707, 2.61)	0.374	0.725	Cases: 34 Controls: 105
- Obesity (>= 30)	1.31 (0.652, 2.66)	0.449	0.738	Cases: 24 Controls: 75
Physical activity at work				
- Sedentary	1 [Reference]			Cases: 11 Controls: 35
- Slightly active	0.446 (0.157, 1.23)	0.12	0.517	Cases: 9 Controls: 47
- Moderately active	0.595 (0.24, 1.49)	0.261	0.725	Cases: 15 Controls: 65
- Fairly active	1.01 (0.455, 2.34)	0.986	0.986	Cases: 31 Controls: 86
- Very active	0.941 (0.34, 2.58)	0.905	0.976	Cases: 10 Controls: 31
Current height (cm)				
- 150 or less	1 [Reference]			Cases: 8 Controls: 38
- 151 - 160	1.65 (0.748, 4.04)	0.24	0.725	Cases: 52 Controls: 156
- 161 - 170	1.38 (0.536, 3.79)	0.512	0.749	Cases: 15 Controls: 63
- 171 or more	0.765 (0.038, 5.39)	0.816	0.934	Cases: 1 Controls: 7
Total fiber (g/day)				
- 0 - 14.21	1 [Reference]			Cases: 15 Controls: 65
- 14.22 - 17.89	1.39 (0.65, 3.04)	0.397	0.725	Cases: 22 Controls: 67
- 17.9 - 23.74	1.51 (0.658, 3.51)	0.331	0.725	Cases: 19 Controls: 60

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
- 23.75 - 77.04	1.34 (0.541, 3.35)	0.528	0.749	Cases: 20 Controls: 72
Total ethanol (g/day)				
- 0 - 0.8	1 [Reference]			Cases: 28 Controls: 102
- 0.81 - 5.18	1.01 (0.523, 1.92)	0.983	0.986	Cases: 22 Controls: 85
- 5.19 - 16.26	1.45 (0.712, 2.93)	0.3	0.725	Cases: 19 Controls: 54
- 16.27 - 151.53	1.04 (0.369, 2.69)	0.934	0.978	Cases: 7 Controls: 23
Milk and yogurt (g/day)				
- 0 - 125	1 [Reference]			Cases: 29 Controls: 56
- 125.01 - 225	0.499 (0.239, 1.02)	0.059	0.472	Cases: 17 Controls: 68
- 225.01 - 369	0.581 (0.292, 1.15)	0.118	0.517	Cases: 23 Controls: 78
- 369.01 - 885.42	0.224 (0.083, 0.541)	0.002	0.138	Cases: 7 Controls: 62
Vegetables (g/day)				
- 0 - 99.84	1 [Reference]			Cases: 10 Controls: 48
- 99.85 - 155.5	0.908 (0.357, 2.35)	0.839	0.934	Cases: 13 Controls: 59
- 155.51 - 222.23	1.51 (0.665, 3.63)	0.337	0.725	Cases: 26 Controls: 73
- 222.24 - 969.09	1.41 (0.618, 3.42)	0.424	0.725	Cases: 27 Controls: 84
Alcoholic beverages (g/day)				
- 0 - 6.56	1 [Reference]			Cases: 29 Controls: 108
- 6.57 - 69.83	1.04 (0.534, 2.01)	0.9	0.976	Cases: 20 Controls: 77
- 69.84 - 250.52	1.4 (0.693, 2.79)	0.343	0.725	Cases: 19 Controls: 55
- 250.53 - 994.07	1.2 (0.45, 2.98)	0.701	0.843	Cases: 8 Controls: 24
Red meat (g/day)				
- 0 - 11.17	1 [Reference]			Cases: 28 Controls: 96

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
- 11.18 - 22	0.84 (0.421, 1.65)	0.616	0.787	Cases: 19 Controls: 75
- 22.01 - 35.34	0.977 (0.473, 1.99)	0.95	0.978	Cases: 17 Controls: 69
- 35.35 - 282.57	2.21 (0.909, 5.29)	0.075	0.472	Cases: 12 Controls: 24
Cured and processed meat (g/day)				
- 0 - 20.35	1 [Reference]			Cases: 26 Controls: 96
- 20.36 - 33.53	1.4 (0.73, 2.71)	0.308	0.725	Cases: 26 Controls: 80
- 33.54 - 49.26	1.37 (0.629, 2.93)	0.422	0.725	Cases: 16 Controls: 55
- 49.27 - 186.94	1.38 (0.488, 3.67)	0.531	0.749	Cases: 8 Controls: 33
Dairy and desserts (g/day)				
- 0 - 2.95	1 [Reference]			Cases: 19 Controls: 82
- 2.96 - 9.29	1.59 (0.758, 3.37)	0.218	0.725	Cases: 20 Controls: 60
- 9.3 - 25.31	1.27 (0.594, 2.72)	0.532	0.749	Cases: 17 Controls: 69
- 25.32 - 405.19	2.11 (0.975, 4.61)	0.059	0.472	Cases: 20 Controls: 53

3.1.2.2 Simple logistic regression models: Males

Table 3 presents findings from logistic regression models assessing CRC risk predictors among male participants. Smoking history emerged as a strong and consistent risk factor. Men who had smoked at least 100 cigarettes or 360 grams of tobacco in their lifetime had more than double the odds of developing CRC compared to never-smokers (OR = 2.43; 95% CI: 1.38–4.45; p= 0.003), with the association remaining significant after false discovery rate (FDR) correction (adjusted p= 0.016). Regular smokers also showed significantly increased risk (OR = 2.43; 95% CI: 1.38–4.45; p= 0.003; adjusted p= 0.016).

Early age of smoking initiation (8–15 years) was associated with higher CRC risk compared to never-smokers (OR = 0.406; 95% CI: 0.206–0.778; p= 0.008; adjusted p= 0.036), while no associations were found for older initiation ages. Current smoking status was a particularly strong predictor, with current smokers exhibiting over four times the odds of CRC compared to non-smokers (OR = 4.26; 95% CI: 2.65–6.96; p< 0.001; adjusted p= 0.007).

Further, both former (OR = 0.26; 95% CI: 0.155–0.43) and never-smokers (OR = 0.171; 95% CI: 0.085–0.332) had significantly lower CRC risk than daily smokers (p< 0.001 for both; adjusted p=

0.007). Smoking status as a categorical variable confirmed these findings, with current smokers having markedly increased CRC risk (OR = 5.82; 95% CI: 3.01–11.7; p< 0.001; adjusted p= 0.007), while no significant association was observed for ex-smokers (OR = 1.52; 95% CI: 0.83–2.87; p= 0.186).

A dose-response relationship was observed for cumulative tobacco exposure. Men with 25.23–42.03 pack-years (OR = 3.28; 95% CI: 1.59–6.57; p= 0.001; adjusted p= 0.007) and 42.04–168.12 pack-years (OR = 2.67; 95% CI: 1.37–5.37; p= 0.005; adjusted p= 0.026) were at elevated CRC risk. Those with 11.02–25.22 pack-years also showed increased odds (OR = 2.22; 95% CI: 1.1–4.55; p= 0.027), though the association was not significant after correction (adjusted p= 0.087).

Average lifetime smoking intensity echoed this pattern. Participants with moderate exposure (3653–7305 cigarettes/year) had significantly increased CRC risk (OR = 3.01; 95% CI: 1.6–5.88; p< 0.001; adjusted p= 0.007), while high exposure (9132–29220 cigarettes/year) was also associated with increased risk (OR = 2.08; 95% CI: 1.06–4.19; p= 0.036), though not significant after correction (adjusted p= 0.103).

Smoking duration further reinforced these associations. Those who smoked for 22–40 years (OR = 2.39; 95% CI: 1.28–4.61; p= 0.007; adjusted p= 0.034) and more than 41 years (OR = 4.57; 95% CI: 2.36–9.16; p< 0.001; adjusted p= 0.007) had elevated CRC risk, whereas shorter durations (1–21 years) showed no significant association.

Behavioral readiness to adopt preventive measures was consistently associated with reduced CRC risk. Men who expressed willingness to change their lifestyle (OR = 0.18; 95% CI: 0.083–0.364; p< 0.001; adjusted p= 0.007), increase physical activity if insufficiently active (OR = 0.154; 95% CI: 0.08–0.283; p< 0.001; adjusted p= 0.007), or reduce meat intake if consuming excessive amounts (OR = 0.226; 95% CI: 0.131–0.384; p< 0.001; adjusted p= 0.007) had significantly lower odds of CRC. Interestingly, those who indicated a willingness to quit smoking had higher odds of CRC (OR = 1.76; 95% CI: 1.15–2.68; p= 0.009; adjusted p= 0.036), possibly reflecting reverse causation or heightened risk perception.

Annual average cigarette consumption during smoking years also correlated with CRC risk. The highest exposure group (7305–29220 cigarettes/year) had significantly increased odds (OR = 3.19; 95% CI: 1.74–6.07; p< 0.001; adjusted p= 0.007), with elevated but non-significant risk in the moderate exposure group (3652–7305 cigarettes/year; OR = 2.29; 95% CI: 1.16–4.63; p= 0.018; adjusted p= 0.06).

Physical activity was inversely associated with CRC risk. Men in the highest METs/week category (31.01–140) had significantly lower odds (OR = 0.419; 95% CI: 0.238–0.732; p= 0.002; adjusted p= 0.013), and a similar trend was observed for moderate activity (17.41–31 METs/week; OR = 0.467; 95% CI: 0.248–0.868; p= 0.017), though the latter did not reach significance after correction (adjusted p= 0.06).

Regarding diet, higher nut consumption was inversely associated with CRC risk (17.15–200 g/day; OR = 0.422; 95% CI: 0.218–0.805; p= 0.009; adjusted p= 0.036). Milk and yogurt intake also showed a protective trend (OR = 0.512; 95% CI: 0.279–0.929; p= 0.029), but did not remain significant after FDR adjustment (adjusted p= 0.089). Similarly, high dietary fiber intake

(23.75–77.04 g/day) was associated with reduced CRC odds (OR = 0.413; 95% CI: 0.201–0.832; p= 0.014), though this did not reach statistical significance post-adjustment (adjusted p= 0.051).

Finally, higher intake of cured and processed meats was positively associated with CRC onset. Participants consuming 33.54–49.26 g/day had elevated risk (OR = 2.09; 95% CI: 1.08–4.13; p= 0.031), though this did not remain statistically significant after correction (adjusted p= 0.092). No significant associations were observed for other intake categories.

Table 3. Simple logistic regression models results for the male subgroup.

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
In your lifetime, have you ever smoked? 'YES' means at least 100 cigarettes or 360 grams of tobacco in your lifetime.				
- No	1 [Reference]			Cases: 18 Controls: 62
- Yes	2.43 (1.38, 4.45)	0.003	0.016	Cases: 130 Controls: 199
Have you ever smoked regularly, i.e., at least one cigarette per day for six months or more?				
- No	1 [Reference]			Cases: 18 Controls: 62
- Yes	2.43 (1.38, 4.45)	0.003	0.016	Cases: 130 Controls: 199
Age at smoking initiation				
- 8 - 15	1 [Reference]			Cases: 44 Controls: 64
- 15 - 17	1.26 (0.694, 2.3)	0.445	0.612	Cases: 36 Controls: 45
- 17 - 19	0.735 (0.385, 1.39)	0.345	0.531	Cases: 24 Controls: 47
- 19 - 54	0.973 (0.51, 1.84)	0.933	0.958	Cases: 26 Controls: 43
- Never smoked	0.406 (0.206, 0.778)	0.008	0.036	Cases: 18 Controls: 62
Current smoker				
- No	1 [Reference]			Cases: 82 Controls: 210
- Yes	4.26 (2.65, 6.96)	p< 0.001	0.007	Cases: 66 Controls: 51
Current frequency of smoking				
- Day	1 [Reference]			Cases: 64 Controls: 50
- Week	0.901 (0.081, 20.2)	0.934	0.958	Cases: 2 Controls: 1
- Former smoker	0.26 (0.155, 0.43)	p< 0.001	0.007	Cases: 64 Controls: 148
- Never	0.171 (0.085, 0.332)	p< 0.001	0.007	Cases: 18 Controls: 62

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
Smoking status				
- Never	1 [Reference]			Cases: 18 Controls: 62
- Former	1.52 (0.83, 2.87)	0.186	0.355	Cases: 64 Controls: 148
- Current	5.82 (3.01, 11.7)	p< 0.001	0.007	Cases: 66 Controls: 51
Pack years				
- Never smoked	1 [Reference]			Cases: 18 Controls: 62
- 0.09 - 11.01	1.81 (0.889, 3.76)	0.105	0.231	Cases: 26 Controls: 55
- 11.02 - 25.22	2.22 (1.1, 4.55)	0.027	0.087	Cases: 30 Controls: 49
- 25.23 - 42.03	3.18 (1.59, 6.57)	0.001	0.007	Cases: 35 Controls: 43
- 42.04 - 168.12	2.67 (1.37, 5.37)	0.005	0.026	Cases: 39 Controls: 52
Average lifetime intensity in cigarettes/year				
- Never smoked	1 [Reference]			Cases: 18 Controls: 62
- 36 - 3652	2.34 (1.21, 4.65)	0.013	0.05	Cases: 39 Controls: 60
- 3653 - 7305	3.01 (1.6, 5.88)	p< 0.001	0.007	Cases: 55 Controls: 69
- 7306 - 9131	0.888 (0.126, 3.94)	0.887	0.949	Cases: 2 Controls: 9
- 9132 - 29220	2.08 (1.06, 4.19)	0.036	0.103	Cases: 34 Controls: 61
Years of smoking				
- Never smoked	1 [Reference]			Cases: 18 Controls: 62
- 1–21 years	0.87 (0.394, 1.9)	0.727	0.848	Cases: 15 Controls: 61
- 22–40 years	2.39 (1.28, 4.61)	0.007	0.034	Cases: 60 Controls: 98
- 41+ years	4.57 (2.36, 9.16)	p< 0.001	0.007	Cases: 55 Controls: 40
Willing to change the lifestyle to reduce colon cancer risk				

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
- No	1 [Reference]			Cases: 32 Controls: 11
- Yes	0.18 (0.083, 0.364)	p< 0.001	0.007	Cases: 116 Controls: 250
If you were a smoker, would you quit smoking?				
- No	1 [Reference]			Cases: 78 Controls: 167
- Yes	1.76 (1.15, 2.68)	0.009	0.036	Cases: 70 Controls: 94
If you were a heavy drinker, would you reduce your alcohol consumption?				
- No	1 [Reference]			Cases: 73 Controls: 141
- Yes	1.25 (0.827, 1.89)	0.291	0.48	Cases: 75 Controls: 120
If you did little exercise: would you do more exercise on a regular basis?				
- No	1 [Reference]			Cases: 46 Controls: 16
- Yes	0.154 (0.08, 0.283)	p< 0.001	0.007	Cases: 102 Controls: 245
If you were to eat a meat-heavy diet: would you eat less meat?				
- No	1 [Reference]			Cases: 50 Controls: 26
- Yes	0.226 (0.131, 0.384)	p< 0.001	0.007	Cases: 98 Controls: 235
Anti-inflammatory medication				
- No	1 [Reference]			Cases: 115 Controls: 202
- Yes	1.08 (0.651, 1.76)	0.772	0.862	Cases: 33 Controls: 59
Passive smoker				
- No	1 [Reference]			Cases: 106 Controls: 201
- Yes	1.36 (0.85, 2.18)	0.195	0.358	Cases: 42 Controls: 60
Average annual cigarettes during the time smoked				

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
- Never smoked	1 [Reference]			Cases: 18 Controls: 62
- 0 - 3652	1.19 (0.541, 2.63)	0.657	0.778	Cases: 16 Controls: 47
- 3652 - 7305	2.29 (1.16, 4.63)	0.018	0.06	Cases: 35 Controls: 55
- 7305 - 29220	3.19 (1.74, 6.07)	p< 0.001	0.007	Cases: 79 Controls: 97
BMI				
- Normal weight (18.5 - 24.9)	1 [Reference]			Cases: 31 Controls: 44
- Overweight (25 - 29.9)	0.696 (0.397, 1.23)	0.207	0.371	Cases: 66 Controls: 124
- Obesity (>= 30)	0.786 (0.439, 1.41)	0.42	0.599	Cases: 51 Controls: 93
Waist-hip ratio >= 1.0				
- < 1	1 [Reference]			Cases: 90 Controls: 176
- >= 1	1.26 (0.818, 1.93)	0.293	0.48	Cases: 58 Controls: 85
Physical activity at work				
- Sedentary	1 [Reference]			Cases: 17 Controls: 35
- Slightly active	0.967 (0.428, 2.19)	0.936	0.958	Cases: 18 Controls: 41
- Moderately active	1.39 (0.698, 2.82)	0.358	0.541	Cases: 50 Controls: 71
- Fairly active	0.821 (0.408, 1.69)	0.586	0.705	Cases: 36 Controls: 87
- Very active	2 (0.906, 4.51)	0.089	0.202	Cases: 27 Controls: 27
METs hours per week				
- 0	1 [Reference]			Cases: 45 Controls: 49
- 0.01 - 17.4	0.545 (0.295, 0.997)	0.05	0.138	Cases: 31 Controls: 59
- 17.41 - 31	0.467 (0.248, 0.868)	0.017	0.06	Cases: 29 Controls: 58
- 31.01 - 140	0.419 (0.238, 0.732)	0.002	0.013	Cases: 43 Controls: 95
METs hours per week walking				

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
- 0	1 [Reference]			Cases: 75 Controls: 126
- 0.01 - 18	0.823 (0.48, 1.39)	0.473	0.616	Cases: 31 Controls: 59
- 18.01 - 108	0.769 (0.464, 1.26)	0.301	0.483	Cases: 42 Controls: 76
Current height (cm)				
- 151 - 160	1 [Reference]			Cases: 8 Controls: 13
- 161 - 170	1.41 (0.556, 3.79)	0.477	0.616	Cases: 74 Controls: 104
- 171 or more	1.03 (0.402, 2.81)	0.946	0.958	Cases: 66 Controls: 144
Total ethanol (g/day)				
- 0 - 0.8	1 [Reference]			Cases: 15 Controls: 47
- 0.81 - 5.18	1.39 (0.653, 3.04)	0.395	0.585	Cases: 25 Controls: 51
- 5.19 - 16.26	1.71 (0.855, 3.55)	0.137	0.27	Cases: 43 Controls: 71
- 16.27 - 151.53	1.95 (1.01, 3.93)	0.054	0.143	Cases: 65 Controls: 92
Legumes (g/day)				
- 0 - 25.33	1 [Reference]			Cases: 32 Controls: 67
- 25.34 - 36.65	0.893 (0.457, 1.73)	0.738	0.848	Cases: 22 Controls: 52
- 36.66 - 43.13	1.08 (0.615, 1.9)	0.795	0.875	Cases: 47 Controls: 87
- 43.14 - 280.95	1.76 (0.97, 3.21)	0.065	0.167	Cases: 47 Controls: 55
Nuts (g/day)				
- 0 - 0.98	1 [Reference]			Cases: 39 Controls: 56
- 0.99 - 6.43	0.994 (0.574, 1.72)	0.981	0.981	Cases: 57 Controls: 80
- 6.44 - 17.14	0.554 (0.287, 1.05)	0.074	0.178	Cases: 26 Controls: 57
- 17.15 - 200	0.422 (0.218, 0.805)	0.009	0.036	Cases: 26 Controls: 68
Alcoholic beverages (g/day)				

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
- 0 - 6.56	1 [Reference]			Cases: 20 Controls: 52
- 6.57 - 69.83	1.27 (0.615, 2.64)	0.52	0.646	Cases: 24 Controls: 47
- 69.84 - 250.52	1.38 (0.721, 2.69)	0.337	0.53	Cases: 41 Controls: 70
- 250.53 - 994.07	1.63 (0.882, 3.08)	0.125	0.26	Cases: 63 Controls: 92
Milk and yogurt (g/day)				
- 0 - 125	1 [Reference]			Cases: 45 Controls: 63
- 125.01 - 225	0.786 (0.442, 1.39)	0.411	0.597	Cases: 37 Controls: 62
- 225.01 - 369	0.696 (0.384, 1.25)	0.23	0.402	Cases: 34 Controls: 59
- 369.01 - 885.42	0.512 (0.279, 0.929)	0.029	0.089	Cases: 32 Controls: 77
Vegetables (g/day)				
- 0 - 99.84	1 [Reference]			Cases: 50 Controls: 80
- 99.85 - 155.5	0.925 (0.545, 1.56)	0.77	0.862	Cases: 44 Controls: 71
- 155.51 - 222.23	0.821 (0.459, 1.46)	0.502	0.634	Cases: 32 Controls: 56
- 222.24 - 969.09	0.572 (0.301, 1.07)	0.083	0.194	Cases: 22 Controls: 54
Total fiber (g/day)				
- 0 - 14.21	1 [Reference]			Cases: 39 Controls: 69
- 14.22 - 17.89	1.06 (0.584, 1.93)	0.843	0.914	Cases: 42 Controls: 56
- 17.9 - 23.74	0.825 (0.447, 1.52)	0.535	0.654	Cases: 42 Controls: 66
- 23.75 - 77.04	0.413 (0.201, 0.832)	0.014	0.051	Cases: 25 Controls: 70
Red meat (g/day)				
- 0 - 11.17	1 [Reference]			Cases: 21 Controls: 43
- 11.18 - 22	1.31 (0.664, 2.62)	0.439	0.612	Cases: 34 Controls: 59

Variable	OR (95% CI)	p-value	Adjusted p-value (FDR)	Counts
- 22.01 - 35.34	1.47 (0.755, 2.91)	0.261	0.447	Cases: 40 Controls: 61
- 35.35 - 282.57	1.27 (0.676, 2.44)	0.461	0.616	Cases: 53 Controls: 98
Cured and processed meat (g/day)				
- 0 - 20.35	1 [Reference]			Cases: 19 Controls: 47
- 20.36 - 33.53	1.3 (0.634, 2.68)	0.48	0.616	Cases: 27 Controls: 54
- 33.54 - 49.26	2.09 (1.08, 4.13)	0.031	0.092	Cases: 50 Controls: 66
- 49.27 - 186.94	1.74 (0.877, 3.56)	0.119	0.255	Cases: 52 Controls: 94
Dairy and desserts (g/day)				
- 0 - 2.95	1 [Reference]			Cases: 26 Controls: 65
- 2.96 - 9.29	1.59 (0.867, 2.96)	0.136	0.27	Cases: 41 Controls: 62
- 9.3 - 25.31	1.79 (0.96, 3.4)	0.069	0.171	Cases: 40 Controls: 61
- 25.32 - 405.19	1.54 (0.814, 2.94)	0.189	0.355	Cases: 41 Controls: 73

3.1.2.3 Multivariate logistic regression models: Females

Table 4 presents the multivariable logistic regression results for female participants. Education level was significantly associated with CRC risk. Women who had completed vocational training (FP or similar) had markedly higher odds of developing CRC compared to those with university-level education (OR = 6.80; 95% CI: 1.70–46.80; *p*= 0.018). Diabetes also emerged as a significant predictor, with diabetic women showing increased odds of CRC relative to non-diabetic counterparts (OR = 2.40; 95% CI: 1.10–5.10; *p*= 0.023). Additionally, older age was associated with elevated CRC risk. Specifically, women aged 61–65 (OR = 2.50; 95% CI: 1.10–5.80; *p*= 0.031) and those aged 66–70 (OR = 3.00; 95% CI: 1.10–7.90; *p*= 0.027) had significantly higher odds of CRC compared to women under 55.

Table 4. Multivariate logistic regression model results for the subgroup of females.

Variables	OR (95% CI)	p-value	β	Counts
METs hours per week 0	1 [Reference]			
0.01 - 17.4	0.5 (0.2-1.1)	0.101	-0.692	Cases: 12 Controls: 82

Variables	OR (95% CI)	p-value	β	Counts
17.41 - 31	0.9 (0.4-1.8)	0.7	-0.149	Cases: 18 Controls: 67
31.01 - 140	1.5 (0.7-3.1)	0.322	0.382	Cases: 22 Controls: 43
Education level University	1 [Reference]			
High school diploma (BUP or COU)	4.5 (1-31.4)	0.072	1.496	Cases: 13 Controls: 40
Vocational training (FP, or similar)	6.8 (1.7-46.8)	0.018	1.923	Cases: 19 Controls: 50
Complete primary education (EGB, or similar)	3.4 (0.9-22.9)	0.117	1.236	Cases: 31 Controls: 117
Incomplete primary education	3.2 (0.6-25.2)	0.194	1.177	Cases: 8 Controls: 19
No formal education, but can read	5.5 (0.6-61.7)	0.139	1.701	Cases: 3 Controls: 4
Waist-hip ratio < 1	1 [Reference]			
Waist-hip ratio >= 1	1.9 (0.6-5.6)	0.229	0.657	Cases: 8 Controls: 10
Diabetes No				
Diabetes Yes	2.4 (1.1-5.1)	0.023	0.876	Cases: 18 Controls: 28
Smoking status Never	1 [Reference]			
Smoking status Former	1.3 (0.6-2.5)	0.499	0.234	Cases: 22 Controls: 78
Smoking status Current	1.5 (0.7-3.3)	0.306	0.408	Cases: 16 Controls: 45
Total energy (kcal/day)	1 (1-1)	0.9	0	
Milk and yogurt (g/day) Below median (< 224)	1 [Reference]			
Milk and yogurt (g/day) Above median (>= 224)	0.6 (0.3-1)	0.063	-0.562	Cases: 30 Controls: 140
Red meat (g/day) Below median (< 16)	1 [Reference]			
Red meat (g/day) Above median (>= 16)	0.7 (0.4-1.3)	0.274	-0.334	Cases: 37 Controls: 133

Variables	OR (95% CI)	p-value	β	Counts
Cured and processed meat (g/day) Below median (< 27)	1 [Reference]			
Cured and processed meat (g/day) Above median (>= 27)	1.6 (0.9-3)	0.117	0.489	Cases: 42 Controls: 128
Dairy and desserts (g/day) Below median (< 8)	1 [Reference]			
Dairy and desserts (g/day) Above median (>= 8)	0.9 (0.5-1.7)	0.829	-0.065	Cases: 37 Controls: 135
Age at recruitment 49 - 55	1 [Reference]			
56 - 60	1.2 (0.5-2.9)	0.7	0.171	Cases: 15 Controls: 63
61 - 65	2.5 (1.1-5.8)	0.031	0.904	Cases: 29 Controls: 70
66 - 70	3 (1.1-7.9)	0.027	1.088	Cases: 19 Controls: 49

3.1.2.4 Multivariate logistic regression models: Males

Table 5 presents the multivariable logistic regression results for male participants. Smoking status remained a significant predictor of CRC onset. Compared to never-smokers, current smokers had significantly increased odds of developing CRC (OR = 3.20; 95% CI: 1.40–7.60; p= 0.006), whereas no significant association was observed for former smokers (OR = 1.40; 95% CI: 0.70–2.80; p= 0.325).

Willingness to adopt lifestyle changes to reduce CRC risk was associated with lower odds of CRC (OR = 0.30; 95% CI: 0.10–0.80; p= 0.016). Similarly, willingness to reduce meat consumption was inversely associated with CRC risk (OR = 0.30; 95% CI: 0.20–0.60; p< 0.001).

Age at recruitment was positively associated with CRC onset. Compared to men aged 49–55 years, those aged 61–65 had significantly higher odds of CRC (OR = 2.80; 95% CI: 1.40–5.80; p= 0.005), and a similar association was observed among those aged 66–70 (OR = 3.30; 95% CI: 1.60–7.20; p= 0.002). No significant association was found for the 56–60 age group.

Table 5. Multivariate logistic regression model results for the subgroup of males.

Variables	OR (95% CI)	p-value	β	Counts
Smoking status Never	1 [Reference]			
Former	1.4 (0.7-2.8)	0.325	0.337	Cases: 64 Controls: 148
Current	3.2 (1.4-7.6)	0.006	1.17	Cases: 66 Controls: 51
Willing to change the lifestyle to reduce colon cancer risk No	1 [Reference]			
Willing to change the lifestyle to reduce colon cancer risk Yes	0.3 (0.1-0.8)	0.016	-1.119	Cases: 116 Controls: 250
If you were a smoker, would you quit smoking? No	1 [Reference]			
If you were a smoker, would you quit smoking? Yes	1.1 (0.6-2)	0.729	0.101	
If you were a heavy drinker, would you reduce your alcohol consumption? No	1 [Reference]			
If you were a heavy drinker, would you reduce your alcohol consumption? Yes	1.7 (1-2.9)	0.057	0.515	
If you were to eat a meat-heavy diet: would you eat less meat? No	1 [Reference]			
If you were to eat a meat-heavy diet: would you eat less meat? Yes	0.3 (0.2-0.6)	p < 0.001	-1.124	
Waist-hip ratio < 1	1 [Reference]			
Waist-hip ratio >= 1	1.1 (0.7-1.8)	0.641	0.117	Cases: 58 Controls: 85
METs hours per week 0	1 [Reference]			
METs hours per week 0.01 - 17.4	0.8 (0.4-1.7)	0.62	-0.174	Cases: 31 Controls: 59
METs hours per week 17.41 - 31	0.6 (0.3-1.3)	0.193	-0.479	Cases: 29 Controls: 58
METs hours per week 31.01 - 140	0.5 (0.3-1)	0.059	-0.615	Cases: 43 Controls: 95

Variables	OR (95% CI)	p-value	β	Counts
Total energy (kcal/day)	1 (1-1)	0.717	0	
Total fiber (g/day) 0 - 14.21	1 [Reference]			
Total fiber (g/day) 14.22 - 17.89	1.6 (0.8-3.4)	0.178	0.494	Cases: 42 Controls: 56
Total fiber (g/day) 17.9 - 23.74	1.3 (0.6-2.9)	0.475	0.284	Cases: 42 Controls: 66
Total fiber (g/day) 23.75 - 77.04	0.8 (0.3-2.2)	0.663	-0.22	Cases: 25 Controls: 70
Total ethanol (g/day) Below median (< 11)	1 [Reference]			
Total ethanol (g/day) Above median (>= 11)	1.1 (0.6-1.8)	0.795	0.068	Cases: 81 Controls: 124
Legumes (g/day) Below median (< 39)	1 [Reference]			
Legumes (g/day) Above median (>= 39)	1.2 (0.7-2.1)	0.423	0.216	Cases: 94 Controls: 140
Nuts (g/day) Below median (< 6)	1 [Reference]			
Nuts (g/day) Above median (>= 6)	0.7 (0.4-1.1)	0.125	-0.412	Cases: 69 Controls: 155
Milk and yogurt (g/day) Below median (< 225)	1 [Reference]			
Milk and yogurt (g/day) Above median (>= 225)	0.7 (0.4-1.2)	0.198	-0.327	Cases: 72 Controls: 148
Red meat (g/day) Below median (< 28)	1 [Reference]			
Red meat (g/day) Above median (>= 28)	1.3 (0.8-2.2)	0.254	0.29	Cases: 78 Controls: 127
Cured and processed meat (g/day) Below median (< 41)	1 [Reference]			
Cured and processed meat (g/day) Above median (>= 41)	1 (0.6-1.8)	0.866	0.046	Cases: 77 Controls: 128
Age at recruitment	1 [Reference]			

Variables	OR (95% CI)	p-value	β	Counts
49 - 55				
56 - 60	1.5 (0.7-3.1)	0.255	0.41	Cases: 34 Controls: 72
61 - 65	2.8 (1.4-5.8)	0.005	1.028	Cases: 44 Controls: 68
66 - 70	3.3 (1.6-7.2)	0.002	1.2	Cases: 45 Controls: 50

3.1.3 Machine learning models for CRC data

Table 6 presents the performance metrics of predictive models developed to identify CRC onset using the full sample. Unlike logistic regression analyses, which were stratified by gender, the machine learning models were initially trained on the combined dataset to maximize statistical power and enhance predictive accuracy. Nonetheless, gender-stratified models are also reported to account for potential sex-specific risk patterns (**Supplemental Table 1, Supplemental Table 2, Supplemental Table 3, Supplemental Table 4, Supplemental Figure 2, Supplemental Figure 3, Supplemental Figure 4, Supplemental Figure 5, Supplemental Figure 6, Supplemental Figure 7, Supplemental Figure 8, Supplemental Figure 9, Supplemental Figure 10, Supplemental Figure 11**).

Table 6. Performance metrics for the best performing models for CRC.

Learner	Accuracy	AUC	PRAUC	F1	Precision	Recall	Macro F1
naive_bayes	0.77±0.035	0.727±0.068	0.594±0.098	0.567±0.068	0.731±0.05	0.694±0.043	0.705±0.045
lda	0.764±0.056	0.715±0.072	0.566±0.124	0.513±0.127	0.728±0.084	0.667±0.074	0.678±0.08
glmnet	0.754 ±0.053	0.714 ±0.071	0.567 ±0.121	0.491+-0.124	0.712±0.077	0.655±0.071	0.664±0.077
xgboost	0.71±0.054	0.662±0.07	0.496±0.1	0.423±0.111	0.641±0.081	0.609±0.067	0.615±0.072

Feature selection for the machine learning models was guided by optimization of predictive performance. The final set of predictors included sociodemographic (marital status), behavioral (physical activity at work, MET-hours per week, maximum weight, age at maximum weight), and smoking-related variables (age at smoking initiation, smoking status). Clinical variables included laxative use and high cholesterol. The models also incorporated psychological and motivational factors such as the emotional impact of cancer-related thoughts during the previous month and participants' stated intentions to lose weight (if obese), increase physical activity (if sedentary), and reduce meat intake (if following a meat-heavy diet). Dietary intake variables included total protein, carbohydrates, ethanol, white meat, cured and processed meat, total meat, fruits, nuts, milk and yogurt, and caloric beverages.

Among the algorithms evaluated—Naïve Bayes, Linear Discriminant Analysis (LDA), Generalized Linear Model with Elastic Net Regularization (GLMNet), and Extreme Gradient Boosting (XGBoost)—Naïve Bayes achieved the best overall performance. It yielded the highest F1-score (0.567 ± 0.068), indicating a favorable balance between precision and recall, and demonstrated the strongest discriminative capacity, with an area under the receiver operating characteristic curve (AUC) of 0.727 ± 0.068.

To evaluate the predictive performance of the models, a confusion matrix (**Figure 1**), receiver operating characteristic (ROC) curve with corresponding area under the curve (AUC) (**Figure 2**), and precision-recall curve (PRAUC) (**Figure 4**) were plotted. These visualizations reflect the output of the Naïve Bayes model, which demonstrated the highest performance among the algorithms tested.

The confusion matrix in **Figure 1** summarizes classification outcomes for a test sample of 749 individuals. The model predicted 76.8% of cases as negative, closely aligning with the actual proportion of 70.1%, and 23.2% as positive, compared to an observed 29.9%. Although a slight underestimation of positive cases was observed, the Naïve Bayes model maintained the most favorable balance between sensitivity and specificity, indicating strong discriminative performance in identifying CRC status.

Figure 1. Confusion matrix for the best performing model for CRC data.

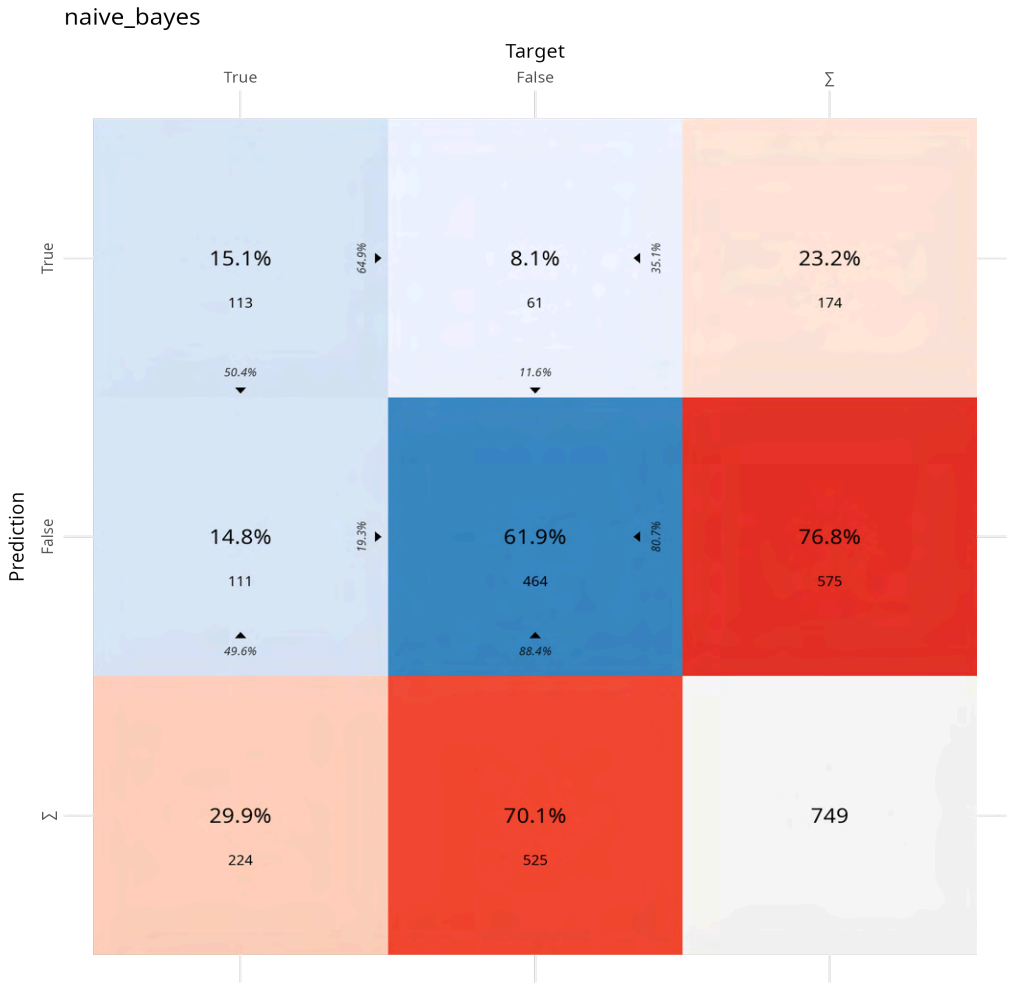

naive_bayes
Target
True
False
Σ
Prediction
True
False
Σ
15.1%
64.9%
113
50.4%
8.1%
35.1%
61
11.6%
23.2%
174
14.8%
19.3%
111
49.6%
61.9%
80.7%
464
88.4%
76.8%
575
29.9%
224
70.1%
525
749

Figure 2 presents the ROC curves illustrating each model's ability to differentiate between CRC and non-CRC cases. Among the models tested, Naïve Bayes achieved the highest area under the curve (AUC) at 0.727, indicating superior discriminative performance. This was followed closely by Linear Discriminant Analysis (LDA) with an AUC of 0.715, and GLM with Elastic Net Regularization (glmnet) with an AUC of 0.714.

Figure 2. Comparison of the area under de ROC curves for CRC models.

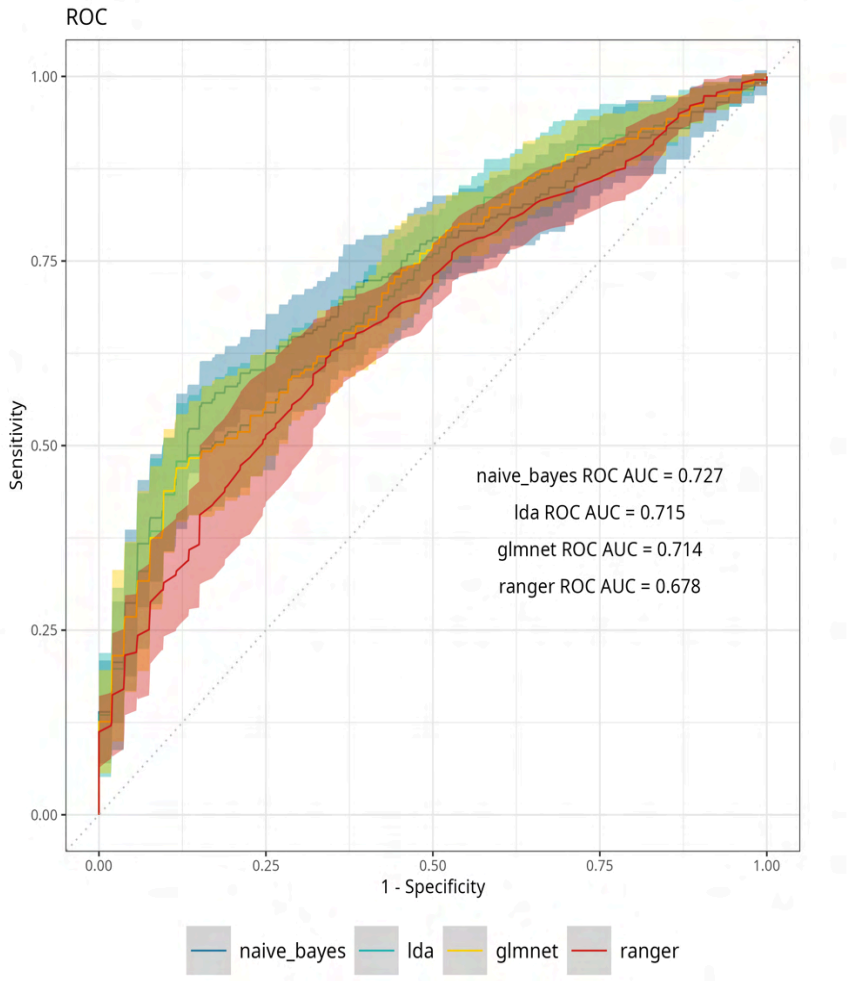


Figure 3 depicts the precision-recall curves, providing an additional measure of model performance. Among the tested models, Naïve Bayes achieved the highest precision-recall area under the curve (PRAUC) at 0.594, followed by GLM with Elastic Net Regularization (glmnet) at 0.567, Linear Discriminant Analysis (LDA) at 0.566, and Random Forest (ranger) at 0.524. These PRAUC values indicate a modest predictive performance—superior to random chance but not yet sufficient for reliable clinical implementation. The results highlight the trade-off between precision (the proportion of true positives among predicted positives) and recall (sensitivity).

Figure 3. Comparison of the area under the precision-Recall curves for CRC models.

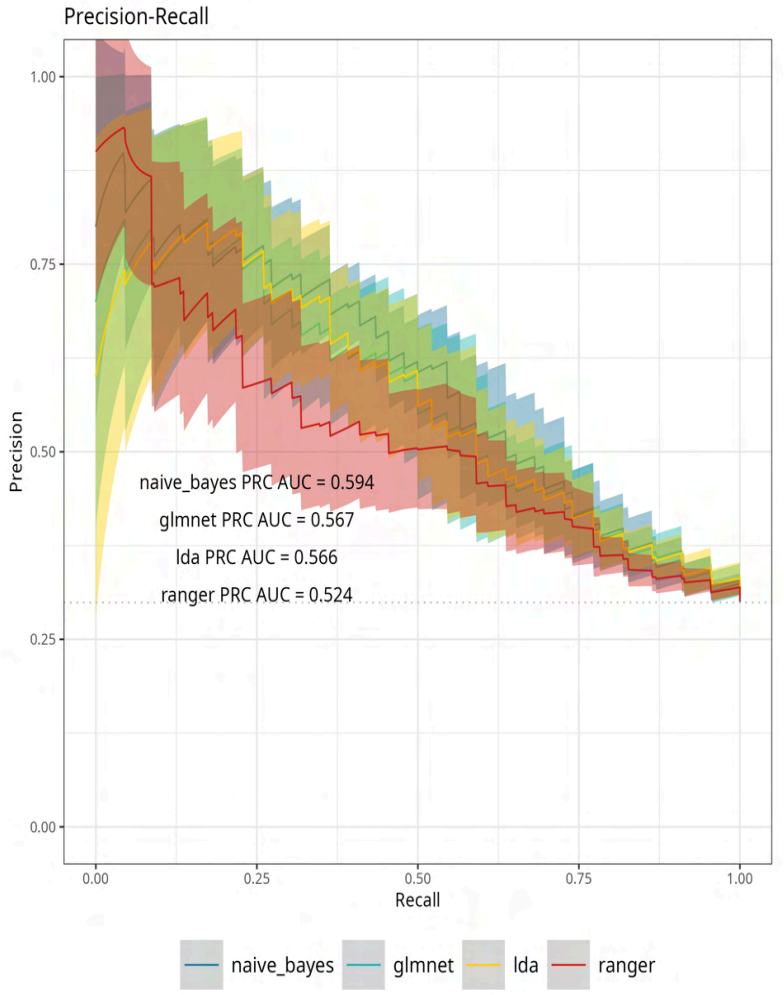
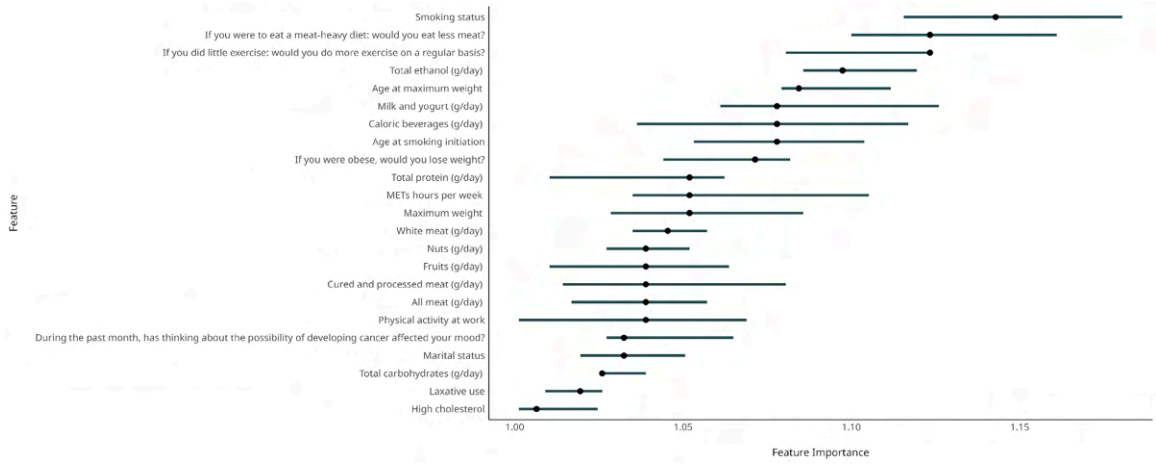


Figure 4 presents the results of the feature importance analysis, indicating that psychological and behavioral variables contributed most significantly to the model's predictive accuracy. Smoking status was identified as the most influential predictor, followed by self-reported willingness to reduce meat consumption and willingness to increase physical activity. Total ethanol consumption (g/day) ranked fourth, underscoring the role of alcohol intake alongside behavioral intentions in shaping CRC risk predictions.

Figure 4. Feature importance for the best prediction model for CRC.

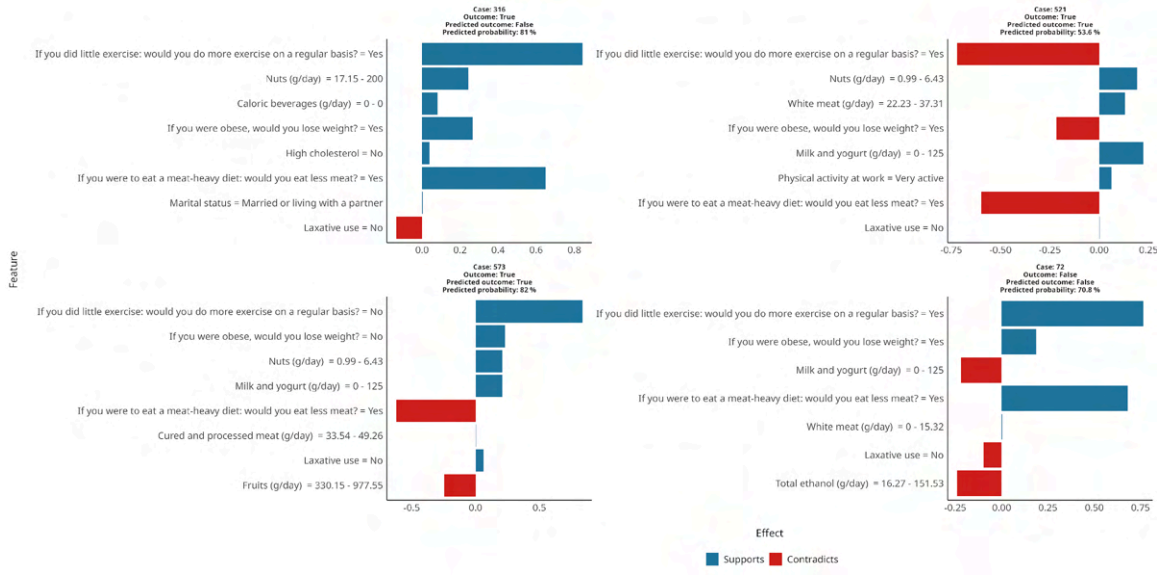


These findings suggest that both concrete behaviors (e.g., smoking and alcohol use) and individuals' readiness to adopt healthier habits are key elements in CRC risk estimation. The high ranking of motivational intent variables indicates that psychological readiness may act as a proxy for latent risk, potentially offering predictive value before the emergence of clinical symptoms. This highlights the utility of incorporating behavioral and psychological dimensions into risk models and supports the development of prevention strategies focused on enhancing health-related motivation and self-efficacy.

Figure 5 displays a Local Interpretable Model-agnostic Explanations (LIME) plot, illustrating how the model generated CRC risk predictions for individual participants. Each bar indicates the influence of a specific variable on the model’s decision, with blue bars supporting the predicted outcome and red bars opposing it.

For example, Case 573 (shown in the bottom left) was classified as having CRC (“True”), with the model assigning an 82% probability to this outcome. Among the factors contributing positively to this prediction, the response to the question “If you did little exercise, would you do more exercise on a regular basis? = No” had the strongest influence. Conversely, the answer “If you were to eat a meat-heavy diet, would you eat less meat? = Yes” worked against the prediction, reflecting the association of this behavior with a reduced CRC risk.

Figure 5. Local Interpretable Model-agnostic Explanations plot for the CRC model.



3.1.3.1 Machine learning models: Female

Supplemental Table 3 summarizes the predictive model performance metrics for the female subgroup. Features included education level, occupation, METs hours per week walking, waist circumference, and BMI. Smoking-related variables comprised smoking status, while health conditions encompassed heartburn, medication for heartburn, diabetes, circulatory problems, and osteoporosis. Reproductive and hormonal factors included age at first menstruation and menopause treatment. Psychological and motivational variables measured the extent of cancer-related worry and its perceived importance, along with willingness to adopt healthier behaviors such as quitting smoking, reducing alcohol intake, and eating less meat. Dietary predictors covered total protein, carbohydrates, fats, fiber, red meat, white meat, cured and processed meat, all meat, fruits, legumes, nuts, dairy and desserts, milk and yogurt, and caloric beverages.

Naïve Bayes was the best-performing algorithm for predicting CRC onset in females, followed by Linear Discriminant Analysis (LDA), Generalized Linear Models with Elastic Net Regularization (GLMNET), and Random Forests (ranger). Naïve Bayes achieved the highest F1-score (0.493 ± 0.132) and area under the curve (AUC) of 0.671 ± 0.116. These modest metrics likely reflect the

small sample size and substantial class imbalance in this subgroup, leading to a decision to focus subsequent machine learning analyses on the combined male and female sample.

Supplemental Figure 2 shows the confusion matrix for the female subgroup's best model. The Naïve Bayes model predicted 73.5% of cases as negative, closely matching the actual negative rate of 77.6%, and 26.5% as positive, compared to an observed 22.4%. Despite some discrepancies, Naïve Bayes demonstrated the most favorable sensitivity-specificity balance among tested models.

Supplemental Figure 3 presents ROC curves, with LDA achieving the highest AUC (0.676), followed by Random Forest (0.674), Naïve Bayes (0.671), and GLMNET (0.657).

Supplemental Figure 4 illustrates precision-recall plots for the female subgroup. LDA had the highest PRAUC (0.45), followed by GLMNET (0.429), Naïve Bayes (0.413), and Random Forest (0.413).

Feature importance analysis in **Supplemental Figure 5** highlights the dominance of psychological and behavioral variables. Willingness to reduce meat consumption was the strongest predictor, followed by milk and yogurt intake and red meat consumption.

The LIME plot in **Supplemental Figure 6** provides participant-level insights into model predictions. For example, Case 279 was correctly classified with a CRC outcome ("True") at a 73.5% predicted probability. Diabetes diagnosis contributed most positively to this prediction, while being employed ("Working") negatively influenced the prediction, consistent with its association with lower cancer risk.

3.1.3.2 Machine learning models: Male

Supplemental Table 4 summarizes the predictive model performance metrics for the male subgroup. Features included education level, marital status, physical activity at work, METs hours per week, waist circumference, current height, maximum weight, and age at maximum weight. Smoking-related predictors comprised age at smoking initiation, passive smoking exposure, and current smoking status. Health condition variables encompassed heartburn, medication for heartburn, diabetes, hypertension, high cholesterol, arthritis, and use of anti-inflammatory medications. Psychological and motivational variables captured the impact of cancer-related thoughts on daily functioning, worry about cancer, and the perceived importance of these concerns. Willingness to adopt healthier behaviors—such as losing weight if obese, quitting smoking, reducing alcohol intake, increasing physical activity, and cutting back on meat—was also assessed. Dietary intake variables included total protein, carbohydrates, fats, fiber, red meat, white meat, all meat, fruits, vegetables, legumes, nuts, dairy and desserts, milk and yogurt, caloric beverages, and alcoholic beverages.

Naïve Bayes was the best-performing algorithm for CRC onset prediction in males, achieving the highest F1-score (0.576 ± 0.087) and demonstrating superior discriminative ability with an area under the curve (AUC) of 0.716 ± 0.088. Other tested models included Neural Networks (nnet), Linear Discriminant Analysis (lda), and GLM with Elastic Net Regularization (glmnet).

Supplemental Figure 7 presents the confusion matrix for the male subgroup's best model. The Naïve Bayes model predicted 68.9% of cases as negative, closely aligning with the actual negative rate of 63.8%, and 31.1% as positive, compared to an observed 36.2%. Despite slight discrepancies, Naïve Bayes exhibited the best balance between sensitivity and specificity.

Supplemental Figure 8 shows ROC curves for the male subgroup. Naïve Bayes achieved the highest AUC (0.716), followed by Neural Networks (0.689), Random Forest (0.682), and Linear Discriminant Analysis (0.662).

Supplemental Figure 9 displays the precision-recall plots. Naïve Bayes had the highest PRAUC (0.625), followed by Random Forest (0.604), Neural Networks (0.582), and k-Nearest Neighbor (0.573). These PRAUC values indicate modest predictive performance—better than chance but not yet optimal for clinical use.

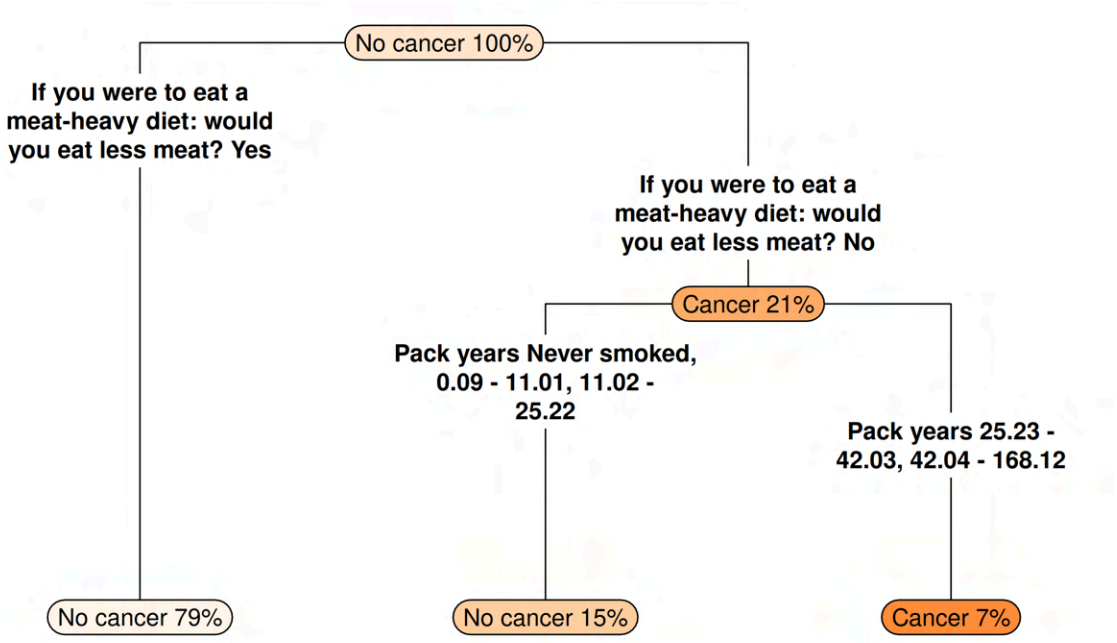
Feature importance analysis (**Supplemental Figure 10**) indicates that psychological and behavioral variables dominated the male subgroup’s model. Smoking status was the most influential predictor, followed by METs hours per week, caloric beverage consumption, and total fats intake.

The LIME plot in **Supplemental Figure 11** provides participant-level interpretability. For instance, Case 308 was correctly classified with CRC (“True”) at a predicted probability of 96.8%. The strongest positive contributor was current smoking status, while willingness to exercise more negatively influenced the prediction, consistent with its association with lower cancer risk.

3.1.4 Regression tree analysis

Figure 6 displays the results of a regression tree analysis, which identifies key factors contributing to CRC risk by iteratively splitting the dataset based on predictive variables. Each node represents a decision point where the model partitions individuals according to the value or category of a specific predictor, ultimately classifying them into terminal nodes reflecting differing levels of CRC risk.

Figure 6. Results of the regression tree analysis for the CRC data.



The analysis incorporated a comprehensive set of variables, including sociodemographic (e.g., education, marital status, occupation), behavioral (e.g., physical activity at work, MET-hours per week, anthropometric measurements), tobacco-related (e.g., age at smoking initiation, passive smoking, current smoking status, pack-years), clinical (e.g., diabetes, hypertension, depression, medication use), psychological (e.g., cancer-related worry, perceived risk, emotional impact), motivational (e.g., willingness to adopt preventive behaviors), and dietary factors (e.g., intake of macronutrients, specific food groups, and total energy).

The strongest predictor identified by the model was participants' willingness to reduce meat consumption in the context of a meat-heavy diet. This variable appeared at the root of the tree, indicating its dominant influence in differentiating CRC risk. Participants who responded "Yes" (indicating openness to dietary change) were routed to the left branch, reaching a terminal node where 79% were classified as CRC-free. This suggests that behavioral readiness to engage in preventive dietary actions may be a protective factor.

Those who answered "No" followed the right branch (21% of the sample), indicating a potentially higher risk group. This subgroup was further stratified by tobacco exposure (pack-years). Participants with no or low smoking exposure (≤25.22 pack-years) were routed left,

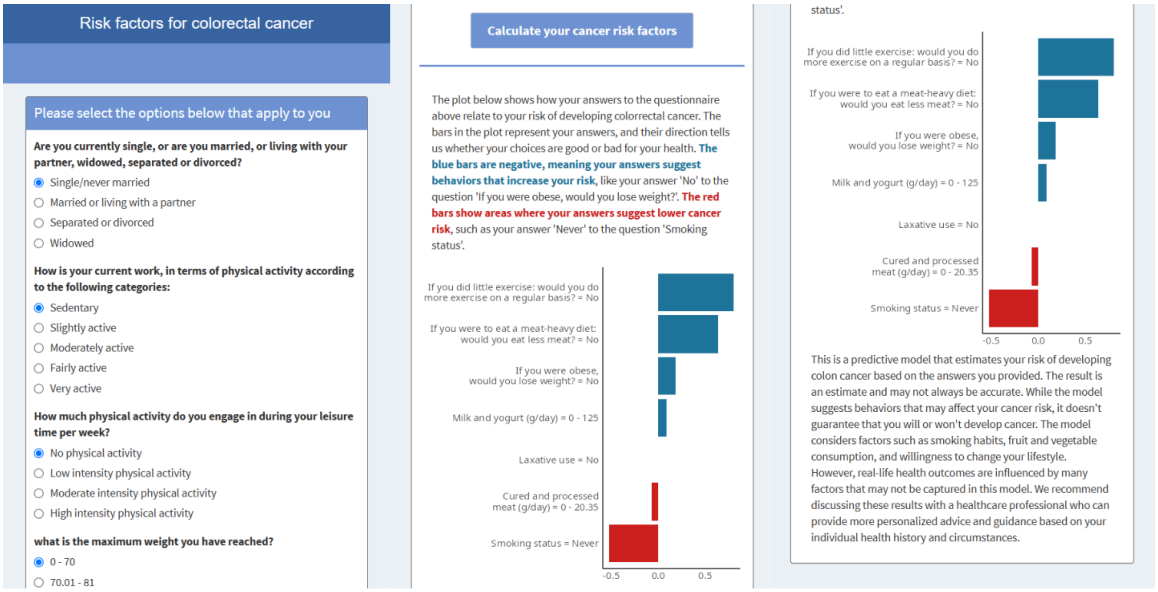
ending in a node where 15% were classified as cancer-free. In contrast, individuals with high cumulative tobacco exposure (>25.23 pack-years) were directed to the right, where only 7% were deemed cancer-free, highlighting a subgroup at elevated risk.

Overall, the tree illustrates the interaction between motivational and behavioral risk factors in shaping CRC outcomes. The lowest predicted risk was observed among those expressing a willingness to modify dietary habits, while the highest risk was found among those resistant to change and with substantial tobacco exposure.

3.1.5 Shiny Application for CRC

A Shiny application was developed to enable end-users to quickly and intuitively evaluate their risk factors for CRC. Upon accessing the application, users are prompted to answer a series of categorical questions corresponding to variables included in the predictive model (Figure 7). After completing the questionnaire, clicking the “Calculate your cancer risk factors” button generates a personalized output displayed as a bar chart, where each bar represents the impact of an individual risk factor. The figure is accompanied by a brief explanation to help users interpret the results, along with a disclaimer clarifying that this is a predictive tool and does not guarantee whether cancer will or will not develop.

Figure 7. Visual Overview of the CRC Risk Prediction Application.



Although the integration of this application into the iBeChange platform was considered, the feasibility is still under evaluation due to the models’ modest predictive performance (e.g., AUC = 0.727). The application is designed to be user-friendly and accessible, with potential plans for translation into Italian, Spanish, and Romanian to broaden accessibility across participating regions.

3.2 COSMOS

3.2.1 Participant characteristics for COSMOS

Table 7 summarizes the characteristics of the 2,690 participants in the COSMOS cohort study, stratified by lung cancer diagnosis. As no statistically significant differences were observed between males and females in the COSMOS cohort, all analyses were conducted using the overall population, in contrast to the gender-stratified approach used in the CRC sample. The cohort comprised 2,580 controls and 110 lung cancer cases. The majority of participants were male (63.4%), with a median age of 62 years [IQR: 58–65].

Unadjusted analyses showed that participants diagnosed with lung cancer were significantly older than controls, with a median age of 63 years (interquartile range [IQR]: 59–68.) compared to 62years [IQR: 58–65] among those without a cancer diagnosis ($p = 0.01$).

Lung cancer participants also reported a longer duration of smoking, with a median of 45 years (IQR: 40–49) compared to 40 years (IQR: 37–45) in the control group ($p < 0.001$). Additionally, the median pack-years was significantly higher among lung cancer cases, at 50 (IQR: 40–70), compared to 40 (IQR: 31–52) in the control group ($p < 0.001$).

Lung-related symptoms were significantly associated with lung cancer in this cohort. Wheezing in the chest was notably more prevalent among lung cancer participants, with 29.4% of them reporting frequent wheezing, compared to 16.2% in the control group ($p < 0.001$). When wheezing occurred, lung cancer cases were more likely to experience it for several days or nights, with 19.8% of lung cancer participants reporting this, compared to 11.0% in the control group ($p = 0.005$). Additionally, shortness of breath during wheezing episodes was significantly more common among lung cancer cases (5.8%) compared to controls (3.9%) ($p = 0.007$). However, breathing normally between wheezing episodes was more common among lung cancer patients, with 12.8% reporting breathing normally between wheezing episodes, compared to 10.2% in the control group ($p = 0.037$).

Lung cancer participants were also more likely to report lung diseases that limited their daily activities in the past year, with 15.9% of lung cancer cases reporting such limitations, compared to 8.2% of controls ($p = 0.009$). Increased production of phlegm during these lung diseases was reported by 12.5% of lung cancer participants, compared to 6.6% of controls ($p = 0.044$). Furthermore, having more than one illness that limited daily activities in the past year was more common among lung cancer participants (4.8%) compared to controls (3.7%) ($p = 0.011$).

Cough was also significantly more prevalent among lung cancer participants, with 62.0% of lung cancer patients reporting a cough, compared to 44.9% in the control group ($p = 0.001$). The daily nature of the cough was more common among lung cancer cases (29.0%) compared to controls (20.7%) ($p < 0.001$). Additionally, intermittent cough was reported more frequently by lung cancer participants (38.2%) than controls (27.3%) ($p = 0.001$).

On the other hand, shortness of breath and phlegm were not significantly associated with lung cancer ($p = 0.138$ and $p = 0.492$, respectively).

A larger proportion of participants with lung cancer also reported being followed by a pulmonologist: 11.6% of lung cancer participants were under the care of a pulmonologist,

compared to 5.9% in the control group (p = 0.043). This is likely related to the higher frequency of lung-related ccomplications among participants who were later diagnosed with lung cancer.

Finally, lung cancer participants exhibited a significantly higher median HADS depression score, with a median score of 4.00 (IQR: 2.00–6.00) among cases, compared to 3.00 (IQR: 1.00–5.00) among controls (p = 0.001).

Supplemental Figure 12 illustrates the distribution of missing data across the dataset, with light blue areas indicating missing responses. The missing data were assumed to be missing at random (MAR) and were handled using multiple imputation techniques. Given the significant class imbalance in the dataset, with only 4% of participants diagnosed with lung cancer, we applied SMOTE (Synthetic Minority Over-sampling Technique) to the data used in the models. Therefore, all subsequent results are based on the dataset after missing data imputation and SMOTE application. **Supplemental Table 5** summarizes the distribution of key variables across the original dataset (N = 2,690), the imputed dataset (N = 2,690), and the class-balanced subset (N = 440). Data are presented separately for individuals with and without lung cancer. This comparative overview provides context for the main analyses and demonstrates the consistency of variable distributions before and after imputation and class balancing.

Table 7. COSMOS study sample characteristics.

Variables	Lung cancer (110)	No cancer (2,580)	p-value	Missing (%)
Female (%)	42 (38.2)	942 (36.5)	0.799	0.0
Age (median [IQR])	63.00 [59.00, 68.00]	62.00 [58.00, 65.00]	0.010	0.0
BMI (median [IQR])	24.62 [22.48, 27.47]	25.38 [22.99, 27.81]	0.158	0.3
Frequency of usual consumption of a portion of raw or cooked vegetables, salad included (150 g) (%)			0.974	2.7
- Rarely (never/1-2 times a month)	5 (4.9)	96 (3.8)		
- Once a week	9 (8.7)	226 (9.0)		
- 2-3 times a week	34 (33.0)	795 (31.6)		
- Every day	41 (39.8)	1063 (42.3)		
- Several times a day	14 (13.6)	335 (13.3)		
Frequency of usual consumption of a portion of fresh fruit (all types - 150 g) (%)			0.324	3.5
- Rarely (never/1-2 times a month)	9 (8.8)	155 (6.2)		
- Once a week	4 (3.9)	186 (7.5)		
- 2-3 times a week	22 (21.6)	547 (21.9)		

Variables	Lung cancer (110)	No cancer (2,580)	p-value	Missing (%)
- Every day	45 (44.1)	1199 (48.1)		
- Several times a day	22 (21.6)	408 (16.4)		
Frequency of usual consumption of a portion of white meat (chicken, turkey, rabbit - 100 g) (%)			0.869	5.1
- Rarely (never/1-2 times a month)	15 (15.3)	379 (15.4)		
- Once a week	39 (39.8)	864 (35.2)		
- 2-3 times a week	42 (42.9)	1145 (46.6)		
- Every day	2 (2.0)	58 (2.4)		
- Several times a day	0 (0.0)	9 (0.4)		
Frequency of usual consumption of a portion of red meat (beef, veal, pork - 100 g) (%)			0.956	6.2
- Rarely (never/1-2 times a month)	18 (17.8)	489 (20.2)		
- Once a week	44 (43.6)	1011 (41.7)		
- 2-3 times a week	38 (37.6)	882 (36.4)		
- Every day	1 (1.0)	39 (1.6)		
- Several times a day	0 (0.0)	1 (0.0)		
Frequency of usual consumption of a portion of cold cuts, cured meats, and sausages (e.g., ham, salami, bresaola/dried beef, sausages, etc. - 50 g) (%)			0.205	4.0
- Rarely (never/1-2 times a month)	16 (15.5)	480 (19.4)		
- Once a week	28 (27.2)	865 (34.9)		
- 2-3 times a week	53 (51.5)	1012 (40.8)		
- Every day	6 (5.8)	113 (4.6)		
- Several times a day	0 (0.0)	10 (0.4)		
Alcohol consumption (e.g., glass of wine, beer, liquor) (%)			0.189	2.1
- Never	44 (41.1)	812 (32.1)		
- ≤4 glasses/week	0 (0.0)	39 (1.5)		

Variables	Lung cancer (110)	No cancer (2,580)	p-value	Missing (%)
- 1-2 glasses/day	40 (37.4)	1165 (46.1)		
- 3-5 glasses/day	21 (19.6)	453 (17.9)		
- >5 glasses/day	2 (1.9)	57 (2.3)		
Have you had any chest diagnostic tests performed in the last year? (%)	20 (18.9)	494 (19.5)	0.979	1.7
Chronic bronchitis (%)	29 (28.2)	423 (16.9)	0.005	3.2
Pneumonia (%)	22 (20.2)	387 (15.2)	0.197	1.0
Tuberculosis (%)	1 (0.9)	47 (1.9)	0.738	2.2
Pleurisy (%)	7 (6.5)	114 (4.5)	0.450	1.8
Pneumothorax (%)	1 (0.9)	31 (1.2)	1.000	2.9
Asthma (%)	4 (3.8)	136 (5.4)	0.619	1.9
Other allergies (%)	15 (14.7)	432 (17.4)	0.563	4.1
Cardiovascular diseases (%)	22 (21.0)	424 (17.1)	0.374	4.0
Thyroid diseases (%)	16 (15.8)	296 (12.2)	0.345	5.9
Other comorbidities (%)	22 (20.0)	400 (15.5)	0.256	0.0
Are you currently undergoing drug therapy? (%)	81 (75.0)	1744 (68.4)	0.181	1.2
Family history of lung cancer (%)	33 (33.7)	688 (29.4)	0.430	9.4
Family member with a history of lung cancer (%)			0.545	9.5
- No family history	65 (66.3)	1650 (70.6)		
- Father	16 (16.3)	368 (15.8)		
- Mother	3 (3.1)	81 (3.5)		
- Brother	4 (4.1)	76 (3.3)		
- Sister	3 (3.1)	25 (1.1)		
- Other	7 (7.1)	136 (5.8)		
Do you currently smoke? = No, former smoker (%)	22 (20.2)	578 (22.5)	0.657	0.3
At what age did you start smoking? (median [IQR])	16.50 [15.00, 18.00]	17.00 [15.00, 19.00]	0.554	0.4

Variables	Lung cancer (110)	No cancer (2,580)	p-value	Missing (%)
For how many years did you smoke in total? (median [IQR])	45.00 [40.00, 49.00]	40.00 [37.00, 45.00]	<0.001	0.6
Pack/years (median [IQR])	50.00 [40.00, 70.00]	40.00 [31.00, 52.00]	<0.001	1.4
Type of cigarettes smoked = Unfiltered (%)	2 (1.9)	46 (1.8)	1.000	2.6
Have you ever smoked cigars? = No (%)	86 (81.9)	2065 (86.9)	0.183	7.8
Have you ever smoked pipes? = No (%)	89 (87.3)	2079 (88.6)	0.800	9.0
Have you ever been exposed to secondhand smoke? (%)	92 (88.5)	2264 (90.0)	0.725	2.6
If you have been exposed to secondhand smoke, specify by whom (%)			0.927	5.0
- Spouse/Partner	15 (15.2)	485 (19.7)		
- At Work	34 (34.3)	818 (33.3)		
- Home/Work	3 (3.0)	82 (3.3)		
- Home/Leisure	1 (1.0)	16 (0.7)		
- Leisure	14 (14.1)	370 (15.1)		
- Leisure/Work	5 (5.1)	97 (3.9)		
- Home/Leisure/Work	3 (3.0)	109 (4.4)		
- Others at Home	12 (12.1)	237 (9.6)		
- Not exposed	12 (12.1)	243 (9.9)		
If you have been exposed to secondhand smoke, how many hours per day? (%)			0.196	15.1
- <1	7 (7.7)	323 (14.7)		
- 2-6	40 (44.0)	998 (45.5)		
- >6	32 (35.2)	622 (28.4)		
- Not exposed	12 (13.2)	250 (11.4)		
Have you ever lived in a big city or near one for more than 10 years? (%)	73 (71.6)	1680 (68.4)	0.572	4.9
Have you ever worked with chemicals? (%)	15 (14.3)	289 (12.6)	0.719	10.8

Variables	Lung cancer (110)	No cancer (2,580)	p-value	Missing (%)
Have you ever been exposed to asbestos? (%)	6 (6.1)	118 (5.5)	0.960	16.2
Have you ever been exposed to cadmium? (%)	0 (0.0)	11 (0.5)	1.000	20.4
Have you ever been exposed to chromium? (%)	0 (0.0)	34 (1.6)	0.418	19.5
Have you ever been exposed to beryllium? (%)	0 (0.0)	6 (0.3)	1.000	20.6
Have you ever been exposed to aluminum? (%)	3 (3.2)	36 (1.7)	0.531	19.3
Have you ever been exposed to silicon dust? (%)	1 (1.1)	31 (1.5)	1.000	19.6
Have you ever been exposed to mixed sulfuric acid? (%)	4 (4.2)	37 (1.8)	0.195	19.3
Have you ever been exposed to ether? (%)	3 (3.2)	27 (1.3)	0.289	19.7
Have you ever been exposed to coal? (%)	1 (1.1)	18 (0.9)	1.000	20.1
Have you ever been exposed to nitrogen mustard? (%)	0 (0.0)	4 (0.2)	1.000	20.6
Have you ever had a Pap smear? (%)			0.123	33.6
- Last year	12 (14.0)	379 (22.3)		
- Last 5 years	12 (14.0)	270 (15.9)		
- No	62 (72.1)	1050 (61.8)		
Have you ever had a mammography? (%)			0.360	32.9
- Last year	21 (23.9)	512 (29.8)		
- Last 5 years	15 (17.0)	225 (13.1)		
- No	52 (59.1)	980 (57.1)		
Have you ever had a colonoscopy or sigmoidoscopy? (%)			0.677	31.3
- Last year	12 (13.5)	252 (14.3)		
- Last 5 years	18 (20.2)	419 (23.8)		
- No	59 (66.3)	1087 (61.8)		

Variables	Lung cancer (110)	No cancer (2,580)	p-value	Missing (%)
Have you ever had a urological exam? (%)			0.615	31.2
- Last year	15 (16.9)	296 (16.8)		
- Last 5 years	10 (11.2)	264 (15.0)		
- No	64 (71.9)	1202 (68.2)		
Have you ever had a PSA test? (%)			0.323	31.0
- Last year	26 (28.9)	612 (34.6)		
- Last 5 years	13 (14.4)	297 (16.8)		
- No	51 (56.7)	858 (48.6)		
Have you ever had a cardiological exam? (%)			0.501	27.3
- Last year	33 (36.3)	618 (33.2)		
- Last 5 years	20 (22.0)	514 (27.6)		
- No	38 (41.8)	732 (39.3)		
Have you ever had a dermatological exam? (%)			0.385	31.0
- Last year	12 (13.6)	300 (17.0)		
- Last 5 years	13 (14.8)	330 (18.7)		
- No	63 (71.6)	1139 (64.4)		
Do you often hear wheezing in your chest? (%)	32 (29.4)	409 (16.2)	<0.001	1.9
If you often hear wheezing in your chest, does it occur for several days or nights? (%)			0.005	3.0
- Yes	21 (19.8)	275 (11.0)		
- No	8 (7.5)	113 (4.5)		
- No wheezing	77 (72.6)	2114 (84.5)		
When wheezing occurs, do you also experience shortness of breath? (%)			0.007	3.8
- Yes	6 (5.8)	97 (3.9)		
- No	21 (20.4)	274 (11.0)		

Variables	Lung cancer (110)	No cancer (2,580)	p-value	Missing (%)
- No wheezing	76 (73.8)	2115 (85.1)		
When you have wheezing, do you breathe normally between episodes? (%)			0.037	7.0
- Yes	12 (12.8)	246 (10.2)		
- No	5 (5.3)	44 (1.8)		
- No wheezing	77 (81.9)	2117 (88.0)		
In the past year, have you suffered from lung diseases that have limited your daily activities for more than a week? (%)	17 (15.9)	204 (8.2)	0.009	3.2
If you have suffered from lung diseases that have limited your daily activities for more than a week in the past year, did you have an increased production of phlegm during such illnesses? (%)			0.044	3.8
- Yes	13 (12.5)	165 (6.6)		
- No	3 (2.9)	43 (1.7)		
- No lung disease	88 (84.6)	2276 (91.6)		
If you have suffered from lung diseases that have limited your daily activities for more than a week, have you had more than one illness of this kind in the past year? (%)			0.011	4.4
- Yes	5 (4.8)	92 (3.7)		
- No	10 (9.6)	95 (3.9)		
- No lung disease	89 (85.6)	2280 (92.4)		
Shortness of breath (%)			0.138	5.8
- I stop because I struggle to breathe after 100 meters or after a few minutes of normal walking on flat ground.	3 (2.9)	24 (1.0)		
- I experience shortness of breath only when I walk quickly on flat ground or on a small incline.	22 (21.4)	524 (21.5)		
- I only experience shortness of breath from exertion.	77 (74.8)	1794 (73.8)		

Variables	Lung cancer (110)	No cancer (2,580)	p-value	Missing (%)
- No	1 (1.0)	90 (3.7)		
Do you have a cough? (%)	67 (62.0)	1135 (44.9)	0.001	2.0
If you have a cough, is it daily? (%)			<0.001	15.6
- Yes	27 (29.0)	451 (20.7)		
- No	26 (28.0)	341 (15.7)		
- No cough	40 (43.0)	1386 (63.6)		
If you have a cough, is it intermittent (%)			0.001	17.3
- Yes	34 (38.2)	583 (27.3)		
- No	14 (15.7)	170 (8.0)		
- No cough	41 (46.1)	1383 (64.7)		
Do you currently have phlegm? (%)	53 (48.6)	1128 (44.8)	0.492	2.3
If you have phlegm, is it mainly in the evening? (%)			0.685	12.9
- Yes	8 (8.2)	153 (6.8)		
- No	34 (34.7)	715 (31.8)		
- No phlegm	56 (57.1)	1377 (61.3)		
Peripheral oxygen saturation at rest (SpO2) (median [IQR])	97.00 [96.00, 98.00]	97.00 [96.00, 98.00]	0.582	2.2
Do you take bronchodilators to improve breathing? (%)	11 (11.8)	137 (6.5)	0.076	18.7
Are you already being followed by a pulmonologist? (%)	11 (11.6)	126 (5.9)	0.043	17.4
Fagerstrom test score (median [IQR])	5.00 [4.00, 7.00]	5.00 [3.00, 6.00]	0.094	26.4
Carbon monoxide level (median [IQR])	2.20 [1.30, 3.40]	2.20 [1.30, 3.40]	0.987	63.6
Parts per million (ppm) of carbon monoxide (median [IQR])	13.00 [4.00, 20.00]	13.00 [4.00, 20.50]	0.898	64.2
HADS Anxiety score (median [IQR])	5.00 [2.00, 7.00]	4.00 [2.00, 7.00]	0.667	11.0
HADS Depression score (median [IQR])	4.00 [2.00, 6.00]	3.00 [1.00, 5.00]	0.001	6.3
HADS Depression category (%)			0.184	6.3
- Normal	90 (85.7)	2170 (89.9)		

Variables	Lung cancer (110)	No cancer (2,580)	p-value	Missing (%)
- Borderline abnormal	9 (8.6)	178 (7.4)		
- Abnormal	6 (5.7)	67 (2.8)		
HADS Anxiety category (%)			0.600	11.0
- Normal	72 (75.8)	1751 (76.2)		
- Borderline abnormal	11 (11.6)	319 (13.9)		
- Abnormal	12 (12.6)	228 (9.9)		

3.2.2 Logistic Regression Models for COSMOS

Table 8 summarize the results of the logistic regression models evaluating individual predictors of lung cancer risk. Each predictor was assessed in a separate model, with all models adjusted for age. Results are reported as odds ratios (ORs) with corresponding 95% confidence intervals (CIs) and p-values.

We assessed the association between various respiratory, lifestyle, and psychosocial factors and the likelihood of lung cancer using age-adjusted logistic regression models. Several variables demonstrated statistically significant associations after adjustment for multiple testing using the False Discovery Rate (FDR) method.

Respiratory symptoms and diagnoses were consistently and strongly associated with lung cancer. Participants with chronic bronchitis had more than double the odds of lung cancer compared to those without the condition (OR = 2.35, 95% CI: 1.5–3.69, p<0.001, FDR p=0.005). Those reporting wheezing had markedly higher odds (OR = 5.28, 95% CI: 3.31–8.57), and this association remained significant (FDR p=0.005). Among those reporting wheezing, individuals who experienced symptoms lasting several days or nights had higher odds of lung cancer compared to those without wheezing (OR = 4.01, 95% CI: 2.33–7.04, FDR p=0.005). Even those with shorter episodes had increased odds (OR = 3.49, FDR p=0.01). Similarly, wheezing was associated with lung cancer, whether or not it was accompanied by shortness of breath. Participants with wheezing but no shortness of breath had slightly higher odds (OR = 4.39, FDR p=0.005) than those who reported both symptoms (OR = 3.21, FDR p=0.016).

Cough was also a strong predictor of lung cancer. Participants who reported having a cough had nearly three times the odds of lung cancer compared to those who did not (OR = 2.87, FDR p=0.005). Among individuals who reported a cough, those with daily (OR = 1.98, FDR p=0.016) or intermittent (OR = 1.91, FDR p=0.012) patterns still had elevated odds compared to those without cough. However, those who had a cough that was neither daily nor intermittent had the highest odds of lung cancer (OR = 2.08 for non-daily; OR = 2.27 for non-intermittent).

Several tobacco-related variables were significantly associated with lung cancer. Compared to individuals with lower cumulative smoking exposure (e.g., 2–35 pack-years), those with higher levels had markedly increased odds. In particular, individuals with 45–60 pack-years had over

twice the odds (OR = 2.3, FDR p=0.012), and those with more than 60 pack-years had nearly fivefold increased odds (OR = 4.68, FDR p=0.005). Similar patterns were observed for duration of smoking, with individuals reporting 45–61 years of smoking showing higher odds than those with fewer years. Interestingly, individuals who never smoked cigars had significantly lower odds of lung cancer than those who had (OR = 0.385, FDR p=0.005). Pipe smoking and age of smoking initiation, however, did not show significant associations.

The Fagerström Test for Nicotine Dependence indicated a trend of increasing lung cancer risk with higher dependence scores. Moderate dependence (score = 5) was associated with more than twice the odds of lung cancer (OR = 2.43, FDR p=0.014), while low dependence (scores 3–4) also showed significance (OR = 2.07, FDR p=0.044). High or very high dependence (scores 6–10) showed increased risk (OR = 2.00), but this was not statistically significant after correction (FDR p=0.079). These findings suggest that nicotine dependence level, particularly in the moderate range, may serve as a potential predictor of lung cancer risk, although further investigation is needed to clarify the risk pattern at higher dependence levels.

Among dietary factors, only daily consumption of cold cuts, cured meats, or sausages was significantly associated with lung cancer (OR = 8.22, 95% CI: 2.96–25.8, FDR p=0.005). Other dietary habits, such as the intake of red meat and vegetables, were not significantly related to cancer risk. Additionally, no significant associations were found between body mass index (BMI) and alcohol consumption.

For psychosocial variables, individuals classified as borderline abnormal on the HADS-Anxiety scale had higher odds of lung cancer (OR = 2.52, FDR p=0.008), but no significant associations were found for the HADS-Depression scale. The "borderline abnormal" classification corresponds to mild or possible anxiety (HADS-Anxiety scores between 8 and 10), suggesting that even low to moderate levels of anxiety symptoms may be linked to increased cancer risk. This raises the possibility that anxiety could contribute to lung cancer risk prediction when considered alongside other factors, though further investigation is needed to clarify the nature of this association.[32]

Table 8. Univariable logistic regression models results for COSMOS.

Variable	Estimate (95% CI)	p-value	Adjusted p values (FDR)	Counts
Chronic bronchitis				
- No	[OR] 1 [Reference]			Cases: 108, Controls: 225
- Yes	[OR] 2.35 (1.5, 3.69)	p < 0.001	p = 0.005	Cases: 57, Controls: 50
Do you often hear wheezing in your chest?				
- No	[OR] 1 [Reference]			Cases: 95, Controls: 241

Variable	Estimate (95% CI)	p-value	Adjusted p values (FDR)	Counts
- Yes	[OR] 5.28 (3.31, 8.57)	p < 0.001	p = 0.005	Cases: 70, Controls: 34
If you often hear wheezing in your chest, does it occur for several days or nights?				
- No wheezing	[OR] 1 [Reference]			Cases: 107, Controls: 241
- Yes	[OR] 4.01 (2.33, 7.04)	p < 0.001	p = 0.005	Cases: 43, Controls: 24
- No	[OR] 3.49 (1.53, 8.3)	p = 0.003	p = 0.01	Cases: 15, Controls: 10
When wheezing occurs, do you also experience shortness of breath?				
- No wheezing	[OR] 1 [Reference]			Cases: 105, Controls: 241
- Yes	[OR] 3.21 (1.39, 7.68)	p = 0.007	p = 0.016	Cases: 14, Controls: 10
- No	[OR] 4.39 (2.57, 7.67)	p < 0.001	p = 0.005	Cases: 46, Controls: 24
When you have wheezing, do you breathe normally between episodes?				
- No wheezing	[OR] 1 [Reference]			Cases: 110, Controls: 241
- Yes	[OR] 3.66 (2.22, 6.12)	p < 0.001	p = 0.005	Cases: 50, Controls: 30
- No	[OR] 2.7 (0.699, 11.1)	p = 0.146	p = 0.224	Cases: 5, Controls: 4
In the past year, have you suffered from lung diseases that have limited your daily activities for more than a week?				
- No	[OR] 1 [Reference]			Cases: 135, Controls: 250
- Yes	[OR] 2.19 (1.24, 3.91)	p = 0.007	p = 0.016	Cases: 30, Controls: 25

Variable	Estimate (95% CI)	p-value	Adjusted p values (FDR)	Counts
If you have suffered from lung diseases that have limited your daily activities for more than a week in the past year, did you have an increased production of phlegm during such illnesses?				
- No lung disease	[OR] 1 [Reference]			Cases: 137, Controls: 246
- Yes	[OR] 1.53 (0.814, 2.86)	p = 0.182	p = 0.246	Cases: 21, Controls: 24
- No	[OR] 2.57 (0.805, 8.86)	p = 0.113	p = 0.186	Cases: 7, Controls: 5
Do you have a cough?				
- No	[OR] 1 [Reference]			Cases: 48, Controls: 149
- Yes	[OR] 2.87 (1.91, 4.35)	p < 0.001	p = 0.005	Cases: 117, Controls: 126
If you have a cough, is it daily?				
- No cough	[OR] 1 [Reference]			Cases: 58, Controls: 145
- Yes	[OR] 1.98 (1.2, 3.27)	p = 0.007	p = 0.016	Cases: 45, Controls: 56
- No	[OR] 2.08 (1.32, 3.29)	p = 0.002	p = 0.008	Cases: 62, Controls: 74
If you have a cough, is it intermittent				
- No cough	[OR] 1 [Reference]			Cases: 60, Controls: 148
- Yes	[OR] 1.91 (1.23, 2.96)	p = 0.004	p = 0.012	Cases: 67, Controls: 86
- No	[OR] 2.27 (1.33, 3.88)	p = 0.003	p = 0.01	Cases: 38, Controls: 41
Do you currently smoke?				
- Yes	[OR] 1 [Reference]			Cases: 139, Controls: 215

Variable	Estimate (95% CI)	p-value	Adjusted p values (FDR)	Counts
- No, former smoker	[OR] 0.628 (0.369, 1.04)	p = 0.079	p = 0.14	Cases: 26, Controls: 60
At what age did you start smoking?				
- 6 - 14	[OR] 1 [Reference]			Cases: 66, Controls: 93
- 14 - 17	[OR] 0.752 (0.484, 1.16)	p = 0.202	p = 0.258	Cases: 65, Controls: 120
- 17 - 30	[OR] 0.729 (0.424, 1.24)	p = 0.247	p = 0.307	Cases: 34, Controls: 62
For how many years did you smoke in total?				
- 14 - 40	[OR] 1 [Reference]			Cases: 65, Controls: 140
- 40 - 45	[OR] 1.23 (0.764, 1.97)	p = 0.393	p = 0.464	Cases: 44, Controls: 77
- 45 - 61	[OR] 2.32 (1.35, 4.05)	p = 0.003	p = 0.01	Cases: 56, Controls: 58
Pack/years				
- 2 - 35	[OR] 1 [Reference]			Cases: 28, Controls: 85
- 35 - 45	[OR] 1.02 (0.563, 1.86)	p = 0.942	p = 0.963	Cases: 30, Controls: 89
- 45 - 60	[OR] 2.3 (1.31, 4.09)	p = 0.004	p = 0.012	Cases: 48, Controls: 63
- Above 60	[OR] 4.68 (2.62, 8.56)	p < 0.001	p = 0.005	Cases: 59, Controls: 38
Have you ever smoked cigars?				
- Yes	[OR] 1 [Reference]			Cases: 45, Controls: 35
- No	[OR] 0.385 (0.234, 0.63)	p < 0.001	p = 0.005	Cases: 120, Controls: 240
Have you ever smoked pipes?				

Variable	Estimate (95% CI)	p-value	Adjusted p values (FDR)	Counts
- Yes	[OR] 1 [Reference]			Cases: 20, Controls: 33
- No	[OR] 1 (0.557, 1.84)	p = 0.997	p = 0.997	Cases: 145, Controls: 242
Fagerstrom test score				
- Very low nicotine dependence (0-2)	[OR] 1 [Reference]			Cases: 20, Controls: 62
- Low nicotine dependence (3-4)	[OR] 2.07 (1.13, 3.87)	p = 0.02	p = 0.044	Cases: 53, Controls: 80
- Moderate nicotine dependence (5)	[OR] 2.43 (1.32, 4.59)	p = 0.005	p = 0.014	Cases: 53, Controls: 69
- High or very high nicotine dependence (6-10)	[OR] 2 (1.05, 3.88)	p = 0.038	p = 0.079	Cases: 39, Controls: 64
Alcohol consumption (e.g., glass of wine, beer, liquor)				
- Never	[OR] 1 [Reference]			Cases: 50, Controls: 84
- From ≤4 glasses/week to 1-2 glasses/day	[OR] 1.07 (0.687, 1.67)	p = 0.772	p = 0.846	Cases: 86, Controls: 136
- >3 glasses/day	[OR] 0.877 (0.492, 1.55)	p = 0.652	p = 0.732	Cases: 29, Controls: 55
BMI				
- Underweight (below 18.5)	[OR] 1 [Reference]			Cases: 6, Controls: 3
- Healthy Weight (18.5 to 24.9)	[OR] 0.332 (0.068, 1.3)	p = 0.127	p = 0.201	Cases: 82, Controls: 127
- Overweight (25 to 29.9)	[OR] 0.278 (0.057, 1.09)	p = 0.077	p = 0.14	Cases: 60, Controls: 110
- Obese (30 or greater)	[OR] 0.25 (0.048, 1.07)	p = 0.071	p = 0.14	Cases: 17, Controls: 35
Frequency of usual consumption of a portion of raw or cooked vegetables, salad included (150 g)				

Variable	Estimate (95% CI)	p-value	Adjusted p values (FDR)	Counts
- Rarely (never/1-2 times a month)	[OR] 1 [Reference]			Cases: 5, Controls: 3
- Once a week	[OR] 0.505 (0.095, 2.28)	p = 0.383	p = 0.464	Cases: 23, Controls: 28
- 2-3 times a week	[OR] 0.357 (0.071, 1.51)	p = 0.17	p = 0.237	Cases: 54, Controls: 90
- Every day	[OR] 0.301 (0.06, 1.27)	p = 0.108	p = 0.184	Cases: 61, Controls: 119
- Several times a day	[OR] 0.36 (0.068, 1.62)	p = 0.191	p = 0.251	Cases: 22, Controls: 35
Frequency of usual consumption of a portion of red meat (beef, veal, pork - 100 g)				
- Rarely (never/1-2 times a month)	[OR] 1 [Reference]			Cases: 23, Controls: 39
- Once a week	[OR] 1.03 (0.572, 1.88)	p = 0.926	p = 0.963	Cases: 72, Controls: 120
- ≥2 times/week	[OR] 1.04 (0.574, 1.9)	p = 0.909	p = 0.963	Cases: 70, Controls: 116
Frequency of usual consumption of a portion of cold cuts, cured meats, and sausages (e.g., ham, salami, bresaola/dried beef, sausages, etc. - 50 g)				
- Rarely (never/1-2 times a month)	[OR] 1 [Reference]			Cases: 18, Controls: 43
- Once a week	[OR] 0.837 (0.432, 1.65)	p = 0.601	p = 0.691	Cases: 37, Controls: 103
- 2-3 times a week	[OR] 1.75 (0.958, 3.3)	p = 0.075	p = 0.14	Cases: 90, Controls: 123
- Every day	[OR] 8.22 (2.96, 25.8)	p < 0.001	p = 0.005	Cases: 20, Controls: 6
HADS Anxiety category				

Variable	Estimate (95% CI)	p-value	Adjusted p values (FDR)	Counts
- Normal	[OR] 1 [Reference]			Cases: 117, Controls: 229
- Borderline abnormal	[OR] 2.52 (1.42, 4.52)	p = 0.002	p = 0.008	Cases: 31, Controls: 25
- Abnormal	[OR] 1.64 (0.819, 3.22)	p = 0.157	p = 0.233	Cases: 17, Controls: 21
HADS Depression category				
- Normal (HADS-D < 8)	[OR] 1 [Reference]			Cases: 148, Controls: 257
- Depression or borderline depression (HADS-D >= 8)	[OR] 1.63 (0.808, 3.27)	p = 0.168	p = 0.237	Cases: 17, Controls: 18

3.2.2.1 Multivariable Models for COSMOS

Next, we assessed the relationship between all predictors and lung cancer onset using a multivariable model. To mitigate collinearity issues, several variables were excluded due to their high correlation with other predictors, which could compromise the accuracy and interpretability of the model. Specifically, the following variables were removed: "Do you often hear wheezing in your chest?", "When you experience wheezing, do you breathe normally between episodes?", "If you have had lung diseases that limited your daily activities for more than a week in the past year, did you have increased phlegm production during those illnesses?", "When wheezing occurs, do you also experience shortness of breath?", "Do you have a cough?", and "Have you ever smoked pipes?" By eliminating these variables, we aimed to reduce redundancy and improve the model's reliability.

Table 9 presents the results of the multivariable logistic regression analysis identifying significant predictors of lung cancer in the study population. Respiratory symptoms, specifically wheezing, emerged as strong predictors of lung cancer. Participants who reported experiencing wheezing that lasted several days or nights had over twice the odds of developing lung cancer compared to those who did not report wheezing (OR = 2.2; 95% CI: 1.0–4.6; p = 0.046). Additionally, those who reported wheezing but not for several days or nights also had increased odds (OR = 4.5; 95% CI: 1.7–12.2; p = 0.003) compared to the referent group.

Smoking exposure, as measured by pack-years, demonstrated a dose-response relationship. Compared to individuals with a lower cumulative smoking history (2–35 pack-years), those with 45–60 pack-years had significantly higher odds of lung cancer (OR = 2.3; 95% CI: 1.1–4.8; p = 0.025), while individuals with more than 60 pack-years had nearly five times the odds (OR = 4.8; 95% CI: 2.0–11.7; p < 0.001). Participants who had never smoked cigars showed a reduced risk of lung cancer compared to those who had (OR = 0.4; 95% CI: 0.2–0.7; p = 0.003).

Vegetable consumption was inversely associated with lung cancer. Compared to participants who rarely consumed vegetables (never or 1–2 times per month), those who consumed them once a week had lower odds (OR = 0.1; 95% CI: 0.0–0.6; p = 0.016), as did those who consumed vegetables 2–3 times a week (OR = 0.1; 95% CI: 0.0–0.6; p = 0.013) and daily (OR = 0.1; 95% CI: 0.0–0.7; p = 0.017).

Conversely, high consumption of processed meats was strongly associated with lung cancer risk. Daily consumers of cold cuts, cured meats, and sausages had dramatically increased odds compared to those who rarely consumed these products (OR = 9.3; 95% CI: 2.8–33.6; p < 0.001).

Table 9. Multivariablelogistic regression models results for COSMOS.

Predictor	OR (95% CI)	p-value	β	Counts
Age 50-59	1 (referent)			Cases: 58, Controls: 94
Age 60-69	0.9 (0.5 - 1.6)	p = 0.739	-0.10 2	Cases: 85, Controls: 156
Age 70 and older	1.4 (0.5 - 3.7)	p = 0.486	0.344	Cases: 22, Controls: 25
Chronic bronchitis No	1 (referent)			Cases: 108, Controls: 225
Chronic bronchitis Yes	1.1 (0.6 - 2.1)	p = 0.642	0.139	Cases: 57, Controls: 50
If you often hear wheezing in your chest, does it occur for several days or nights? No wheezing	1 (referent)			Cases: 107, Controls: 241
If you often hear wheezing in your chest, does it occur for several days or nights? Yes	2.2 (1 - 4.6)	p = 0.046	0.766	Cases: 43, Controls: 24
If you often hear wheezing in your chest, does it occur for several days or nights? No	4.5 (1.7 - 12.2)	p = 0.003	1.499	Cases: 15, Controls: 10
In the past year, have you suffered from lung diseases that have limited your daily activities for more than a week? No	1 (referent)			Cases: 135, Controls: 250
In the past year, have you suffered from lung diseases that have limited your daily activities for more than a week? Yes	1.8 (0.9 - 3.6)	p = 0.116	0.571	Cases: 30, Controls: 25
If you have a cough, is it intermittent No cough	1 (referent)			Cases: 60, Controls: 148
If you have a cough, is it intermittent Yes	1.4 (0.8 - 2.4)	p = 0.254	0.326	Cases: 67, Controls: 86
If you have a cough, is it intermittent No	1.6 (0.8 - 3.3)	p = 0.215	0.459	Cases: 38, Controls: 41

Predictor	OR (95% CI)	p-value	β	Counts
Do you currently smoke? Yes	1 (referent)			Cases: 139, Controls: 215
Do you currently smoke? No, former smoker	0.7 (0.3 - 1.5)	p = 0.38	-0.328	Cases: 26, Controls: 60
At what age did you start smoking? 6 - 14	1 (referent)			Cases: 66, Controls: 93
At what age did you start smoking? 14 - 17	1 (0.6 - 1.8)	p = 0.869	0.047	Cases: 65, Controls: 120
At what age did you start smoking? 17 - 30	1.4 (0.7 - 2.9)	p = 0.362	0.339	Cases: 34, Controls: 62
For how many years did you smoke in total? 14 - 40	1 (referent)			Cases: 65, Controls: 140
For how many years did you smoke in total? 40 - 45	1 (0.5 - 1.8)	p = 0.927	-0.03	Cases: 44, Controls: 77
For how many years did you smoke in total? 45 - 61	1.1 (0.5 - 2.5)	p = 0.733	0.133	Cases: 56, Controls: 58
Pack/years 2 - 35	1 (referent)			Cases: 28, Controls: 85
Pack/years 35 - 45	0.9 (0.4 - 1.9)	p = 0.815	-0.085	Cases: 30, Controls: 89
Pack/years 45 - 60	2.3 (1.1 - 4.8)	p = 0.025	0.83	Cases: 48, Controls: 63
Pack/years Above 60	4.8 (2 - 11.7)	p < 0.001	1.568	Cases: 59, Controls: 38
Have you ever smoked cigars? Yes	1 (referent)			Cases: 45, Controls: 35
Have you ever smoked cigars? No	0.4 (0.2 - 0.7)	p = 0.003	-0.929	Cases: 120, Controls: 240
Fagerstrom test score Very low nicotine dependence (0-2)	1 (referent)			Cases: 20, Controls: 62
Fagerstrom test score Low nicotine dependence (3-4)	1.7 (0.8 - 3.6)	p = 0.169	0.518	Cases: 53, Controls: 80
Fagerstrom test score Moderate nicotine dependence (5)	1.3 (0.6 - 2.8)	p = 0.555	0.239	Cases: 53, Controls: 69
Fagerstrom test score High or very high nicotine dependence (6-10)	0.9 (0.3 - 2.1)	p = 0.727	-0.161	Cases: 39, Controls: 64
Alcohol consumption (e.g., glass of wine, beer, liquor) Never	1 (referent)			Cases: 50, Controls: 84

Predictor	OR (95% CI)	p-value	β	Counts
Alcohol consumption (e.g., glass of wine, beer, liquor) From ≤4 glasses/week to 1-2 glasses/day	1 (0.6 - 1.8)	p = 0.892	0.038	Cases: 86, Controls: 136
Alcohol consumption (e.g., glass of wine, beer, liquor) >3 glasses/day	0.8 (0.4 - 1.5)	p = 0.47	-0.256	Cases: 29, Controls: 55
BMI Underweight (below 18.5)	1 (referent)			Cases: 6, Controls: 3
BMI Healthy Weight (18.5 to 24.9)	0.5 (0.1 - 2.4)	p = 0.387	-0.722	Cases: 82, Controls: 127
BMI Overweight (25 to 29.9)	0.3 (0.1 - 1.5)	p = 0.164	-1.186	Cases: 60, Controls: 110
BMI Obese (30 or greater)	0.2 (0 - 1.2)	p = 0.085	-1.542	Cases: 17, Controls: 35
Frequency of usual consumption of a portion of raw or cooked vegetables, salad included (150 g) Rarely (never/1-2 times a month)	1 (referent)			Cases: 5, Controls: 3
Frequency of usual consumption of a portion of raw or cooked vegetables, salad included (150 g) Once a week	0.1 (0 - 0.6)	p = 0.016	-2.201	Cases: 23, Controls: 28
Frequency of usual consumption of a portion of raw or cooked vegetables, salad included (150 g) 2-3 times a week	0.1 (0 - 0.6)	p = 0.013	-2.126	Cases: 54, Controls: 90
Frequency of usual consumption of a portion of raw or cooked vegetables, salad included (150 g) Every day	0.1 (0 - 0.7)	p = 0.017	-2.033	Cases: 61, Controls: 119
Frequency of usual consumption of a portion of raw or cooked vegetables, salad included (150 g) Several times a day	0.2 (0 - 1)	p = 0.055	-1.714	Cases: 22, Controls: 35
Frequency of usual consumption of a portion of red meat (beef, veal, pork - 100 g) Rarely (never/1-2 times a month)	1 (referent)			Cases: 23, Controls: 39
Frequency of usual consumption of a portion of red meat (beef, veal, pork - 100 g) Once a week	1 (0.5 - 2.1)	p = 0.991	-0.004	Cases: 72, Controls: 120
Frequency of usual consumption of a portion of red meat (beef, veal, pork - 100 g) ≥2 times/week	0.8 (0.4 - 1.7)	p = 0.591	-0.205	Cases: 70, Controls: 116
Frequency of usual consumption of a portion of cold cuts, cured meats, and sausages (e.g., ham, salami, bresaola/dried beef, sausages, etc. - 50 g) Rarely (never/1-2 times a month)	1 (referent)			Cases: 18, Controls: 43
Frequency of usual consumption of a portion of cold cuts, cured meats, and sausages (e.g., ham, salami,	0.8 (0.4 - 1.8)	p = 0.645	-0.183	Cases: 37, Controls: 103

Predictor	OR (95% CI)	p-value	β	Counts
bresaola/dried beef, sausages, etc. - 50 g) Once a week				
Frequency of usual consumption of a portion of cold cuts, cured meats, and sausages (e.g., ham, salami, bresaola/dried beef, sausages, etc. - 50 g) 2-3 times a week	1.5 (0.7 - 3.1)	p = 0.32	0.374	Cases: 90, Controls: 123
Frequency of usual consumption of a portion of cold cuts, cured meats, and sausages (e.g., ham, salami, bresaola/dried beef, sausages, etc. - 50 g) Every day	9.3 (2.8 - 33.6)	p < 0.001	2.227	Cases: 20, Controls: 6
HADS Depression category Normal (HADS-D < 8)	1 (referent)			Cases: 148, Controls: 257
HADS Depression category Depression or borderline depression (HADS-D >= 8)	1.4 (0.6 - 3.3)	p = 0.501	0.303	Cases: 17, Controls: 18
HADS Anxiety category Normal	1 (referent)			Cases: 117, Controls: 229
HADS Anxiety category Borderline abnormal	1.7 (0.8 - 3.6)	p = 0.151	0.536	Cases: 31, Controls: 25
HADS Anxiety category Abnormal	1.5 (0.6 - 3.7)	p = 0.345	0.43	Cases: 17, Controls: 21

3.2.3 Machine learning model for COSMOS data

Table 10 presents the performance metrics of the machine learning models used to predict lung cancer risk. The feature selection process for the models was guided by the goal of optimizing predictive performance. The final set of predictors included BMI, frequency of usual consumption of a portion of raw or cooked vegetables (150 g), frequency of usual consumption of a portion of fresh fruit (150 g), frequency of usual consumption of a portion of white meat (100 g), and frequency of usual consumption of a portion of red meat (100 g). Also included were the frequency of usual consumption of cold cuts, cured meats, and sausages (50 g), alcohol consumption (e.g., glass of wine, beer, liquor), and whether the participant had any chest diagnostic tests performed in the last year. Other predictors included current drug therapy, smoking status, total years smoked, pack/years, and history of smoking cigars or pipes. The analysis also considered secondhand smoke exposure, including the source and hours per day, as well as whether the participant had lived in a big city or near one for more than 10 years. Occupational exposure to chemicals, recent medical exams (e.g., Pap smear, mammography, colonoscopy, sigmoidoscopy, urological exam, PSA test, cardiological exam, and dermatological exam), and whether the participant had suffered from lung diseases limiting daily activities in the past year were also included. Additionally, the Fagerstrom test score, HADS Anxiety category, and HADS Depression category were part of the final set of predictors. Categorical variables were transformed into dummy logical variables to enhance model performance.

Among the algorithms tested, including Linear Discriminant Analysis (lda), GLM with Elastic Net Regularization (glmnet), and k-Nearest Neighbors (kknn), Extreme Gradient Boosting (xgboost)

achieved the highest F1-score, 0.540 (±0.081), and demonstrated the best discriminative ability with an AUC of 0.710 (±0.071).

Table 10. Performance metrics for the best performing models for lung cancer.

Learner	Accuracy	AUC	PRAUC	F1	Precision	Recall	Macro F1
xgboost	0.688 (+-0.062)	0.71 (+-0.071)	0.634 (+-0.057)	0.54 (+-0.081)	0.677 (+-0.072)	0.65 (+-0.061)	0.65 (+-0.06)
glmnet	0.67 (+-0.072)	0.667 (+-0.092)	0.586 (+-0.116)	0.517 (+-0.124)	0.653 (+-0.092)	0.634 (+-0.081)	0.631 (+-0.082)
kknn	0.65 (+-0.035)	0.707 (+-0.067)	0.633 (+-0.071)	0.513 (+-0.077)	0.63 (+-0.042)	0.622 (+-0.043)	0.617 (+-0.041)
lda	0.666 (+-0.071)	0.676 (+-0.085)	0.586 (+-0.119)	0.511 (+-0.118)	0.647 (+-0.083)	0.63 (+-0.08)	0.626 (+-0.079)

The confusion matrix in **Figure 8** presents the classification results of the xgboost model on a test set of 440 instances. The model predicted 30.9% of cases as positive, while the actual positive rate was 37.5%. Conversely, 69.1% cases were predicted as negative, compared to an observed negative rate of 62.5%. The largest proportion of correct classifications occurred among true negatives (50.2%), whereas true positives accounted for 18.6% of the predictions. Although there were variations, xgboost achieved the most balanced results between specificity and sensitivity among all models evaluated.

Figure 8. Confusion matrix for the best performing model for COSMOS data.

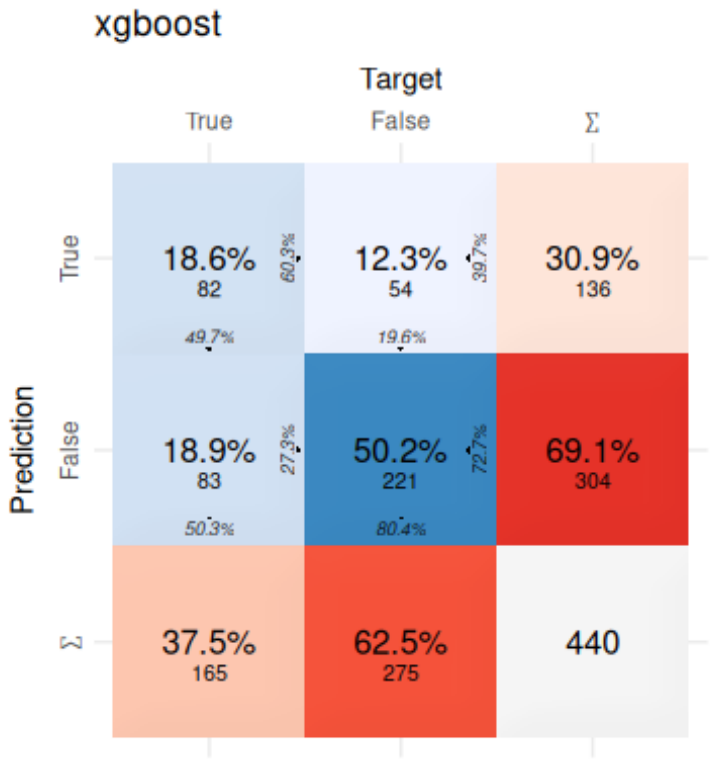


Figure 9 displays the ROC curves used to assess the models’ ability to distinguish between cancer and non-cancer cases. The ranger model achieved the highest AUC (0.715), followed closely by xgboost (0.710) and kknn (0.707). Linear Discriminant Analysis (lda) yielded a lower discriminative performance with an AUC of 0.676.

Figure 9. Comparison of the area under de ROC curves for COSMOS models.

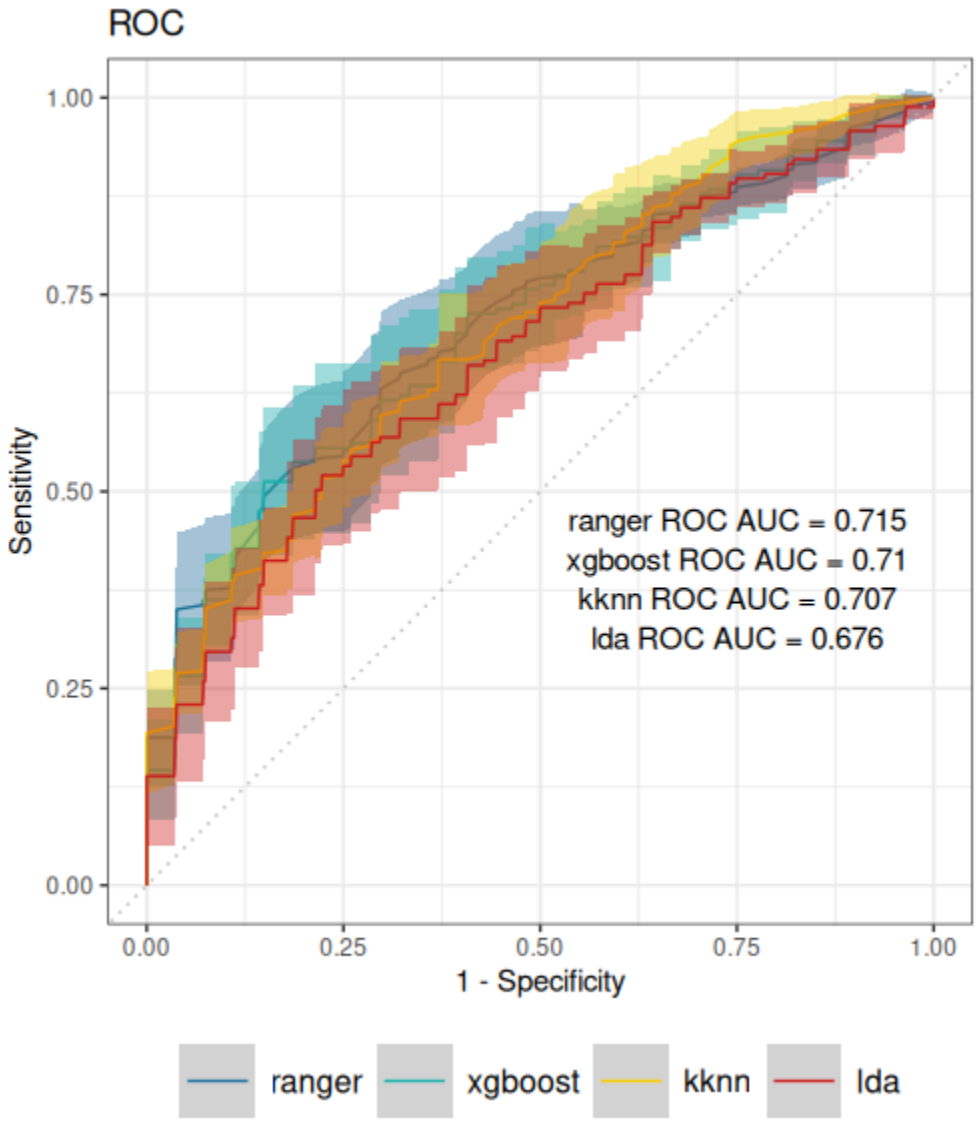
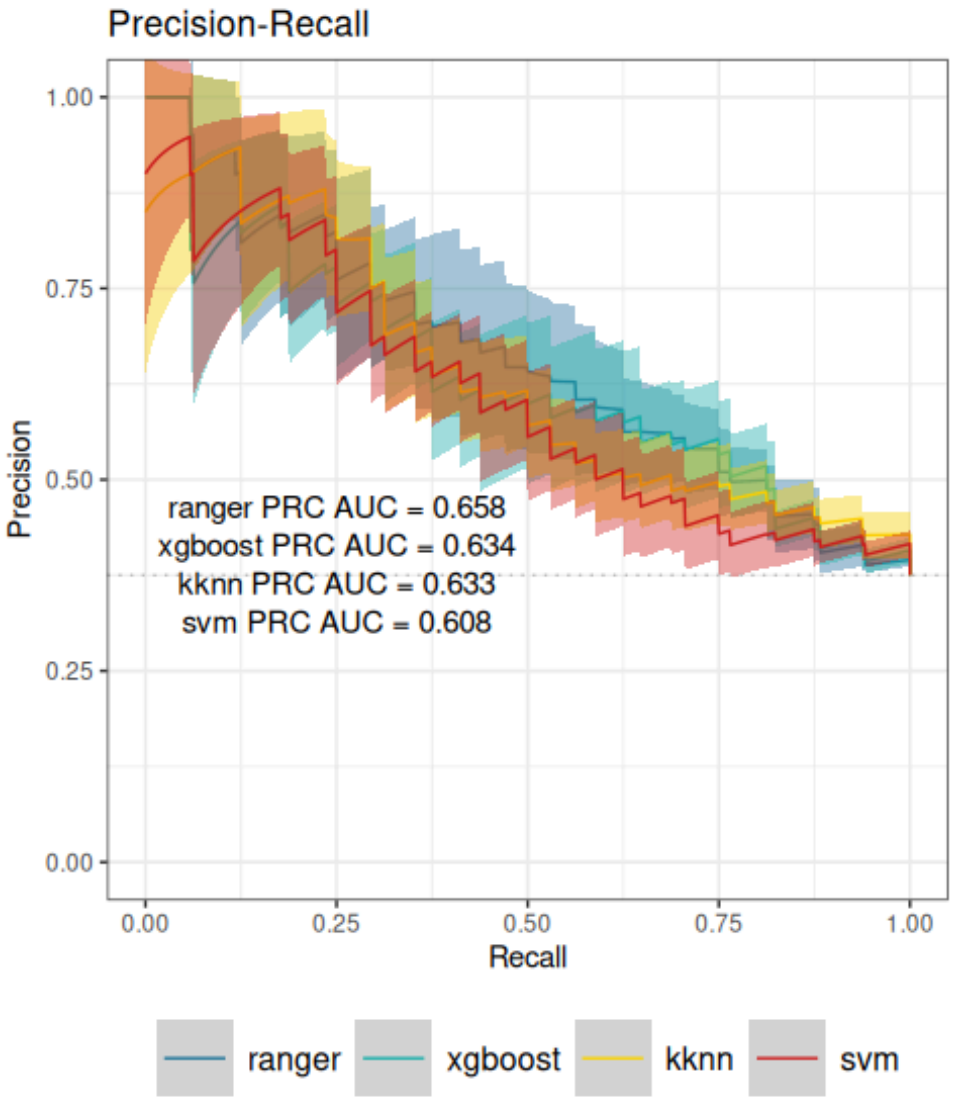


Figure 10 illustrates the precision-recall curves, providing further insight into model performance in the context of class imbalance. Ranger again outperformed the others, with a PRAUC of 0.658, followed by xgboost (0.634), kknn (0.633), and support vector machines (svm, 0.608). These results suggest that ranger and xgboost maintained the best balance between precision and recall (sensitivity) across the tested models, yet values are still not ideal for clinical applications.

Figure 10. Comparison of the area under the precision-Recall curves for COSMOS models.



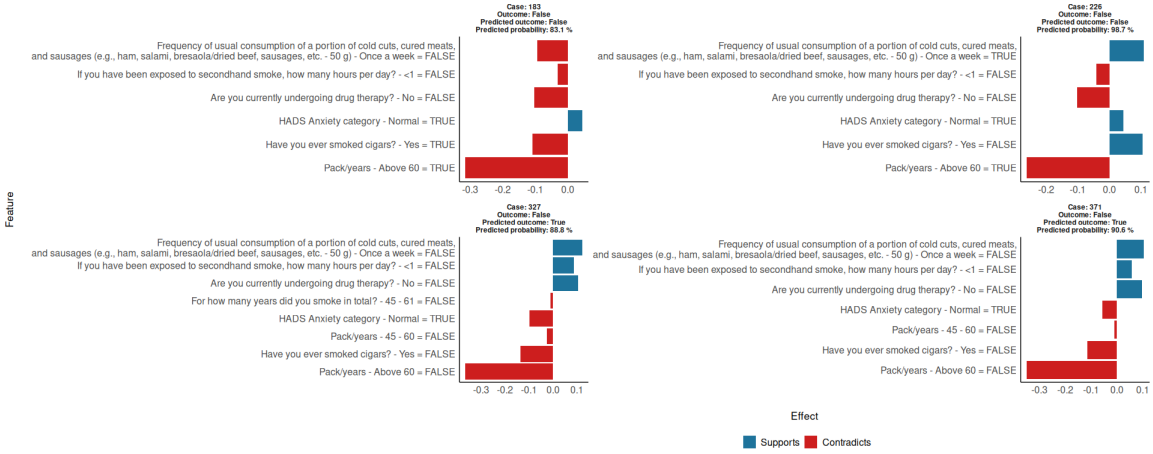
The feature importance analysis, presented in **Figure 11**, indicates that tobacco exposure was by far the strongest influencer in the lung cancer prediction models. The highest-ranking feature was cumulative smoking history (pack/years above 60), followed by absence of current drug therapy. Regular consumption of cold cuts, cured meats, and sausages (50g once a week) also contributed meaningfully to the model's performance. Notably, exposure to secondhand smoke in the workplace emerged as a relevant predictor, underscoring the impact of both direct and environmental tobacco exposure on lung cancer risk estimation. Other variables also influenced the model, though their contributions were comparatively weaker.

Figure 11. Feature importance for the best predictive model for lung cancer.


Pack/years - Above 60
Are you currently undergoing drug therapy? - No
Frequency of usual consumption of a portion of cold cuts, cured meats, and sausages (e.g., ham, salami, bresaola/dried beef, sausages, etc. - 50 g) - Once a week
If you have been exposed to secondhand smoke, specify by whom - At Work
If you have been exposed to secondhand smoke, how many hours per day? - Not exposed
Fagerstrom test score - Moderate nicotine dependence (5)
For how many years did you smoke in total? - 14 - 40
If you have been exposed to secondhand smoke, specify by whom - Spouse/Partner
If you have been exposed to secondhand smoke, how many hours per day? - <1
Have you ever smoked cigars? - Yes
Have you ever had a dermatological exam? - No
Do you currently smoke? - Yes
HADS Anxiety category - Normal
Have you ever worked with chemicals?
Frequency of usual consumption of a portion of white meat (chicken, turkey, rabbit - 100 g) - Rarely (never/1-2 times a month)
For how many years did you smoke in total? - 40 - 45
If you have been exposed to secondhand smoke, specify by whom - Leisure
Have you ever had a PSA test? - Last 5 years
Pack/years - 45 - 60
Have you ever lived in a big city or near one for more than 10 years?
Frequency of usual consumption of a portion of fresh fruit (all types - 150 g) - Every day
Have you ever had a dermatological exam? - Last year
Have you ever had a cardiological exam? - Last 5 years
BMI - Overweight (25 to 29.9)
BMI - Healthy Weight (18.5 to 24.9)
Alcohol consumption (e.g., glass of wine, beer, liquor) - 3-5 glasses/day
Alcohol consumption (e.g., glass of wine, beer, liquor) - 1-2 glasses/day
Frequency of usual consumption of a portion of white meat (chicken, turkey, rabbit - 100 g) - Once a week
Frequency of usual consumption of a portion of white meat (chicken, turkey, rabbit - 100 g) - 2-3 times a week
Frequency of usual consumption of a portion of raw or cooked vegetables, salad included (150 g) - Several times a day
Frequency of usual consumption of a portion of raw or cooked vegetables, salad included (150 g) - Once a week
Frequency of usual consumption of a portion of raw or cooked vegetables, salad included (150 g) - Every day
Frequency of usual consumption of a portion of raw or cooked vegetables, salad included (150 g) - 2-3 times a week
Have you ever had a urological exam? - No
Have you ever had a urological exam? - Last year
Have you ever had a urological exam? - Last 5 years
For how many years did you smoke in total? - 45 - 61
At what age did you start smoking? - 6 - 14
At what age did you start smoking? - 17 - 30
At what age did you start smoking? - 14 - 17
If you have been exposed to secondhand smoke, specify by whom - Others at Home
If you have been exposed to secondhand smoke, how many hours per day? - >6
If you have been exposed to secondhand smoke, how many hours per day? - 2-6
Frequency of usual consumption of a portion of red meat (beef, veal, pork - 100 g) - Rarely (never/1-2 times a month)
Frequency of usual consumption of a portion of red meat (beef, veal, pork - 100 g) - Once a week
Frequency of usual consumption of a portion of red meat (beef, veal, pork - 100 g) - 2-3 times a week
Have you ever had a PSA test? - No
Have you ever had a PSA test? - Last year
Frequency of usual consumption of a portion of cold cuts, cured meats, and sausages (e.g., ham, salami, bresaola/dried beef, sausages, etc. - 50 g) - Rarely (never/1-2 times a month)
Frequency of usual consumption of a portion of cold cuts, cured meats, and sausages (e.g., ham, salami, bresaola/dried beef, sausages, etc. - 50 g) - 2-3 times a week
Recent Pap smear or mammography - Pap smear or mammography in last year
Pack/years - 35 - 45
Pack/years - 2 - 35
Frequency of usual consumption of a portion of fresh fruit (all types - 150 g) - Several times a day
Frequency of usual consumption of a portion of fresh fruit (all types - 150 g) - 2-3 times a week
Fagerstrom test score - Very low nicotine dependence (0-2)
Fagerstrom test score - Low nicotine dependence (3-4)
Fagerstrom test score - High or very high nicotine dependence (6-10)
Have you ever smoked pipes? - Yes
Have you ever had a dermatological exam? - Last 5 years
Have you ever had a colonoscopy or sigmoidoscopy? - No
Have you ever had a colonoscopy or sigmoidoscopy? - Last year
Have you ever had a colonoscopy or sigmoidoscopy? - Last 5 years
Have you had any chest diagnostic tests performed in the last year?
Have you ever had a cardiological exam? - No
Have you ever had a cardiological exam? - Last year
BMI - Obese (30 or greater)
HADS Anxiety category - Borderline abnormal
Alcohol consumption (e.g., glass of wine, beer, liquor) - Never
Feature
0.00
0.02
0.04
0.06
Feature Importance

Figure 12 shows a Local Interpretable Model-agnostic Explanations (LIME) plot, which illustrates how individual features contributed to the lung cancer prediction for individual participants. For instance, in the top left for Case 183, the model classified this participant as not having lung cancer (“False”), with a predicted probability of 83.1%. Each bar reflects the influence of a specific variable on the model’s decision, with blue bars supporting the outcome and red bars contradicting it. The most influential variable contradicting this prediction was “Pack/years – Above 60 = TRUE”, while the variable “HADS Anxiety category - Normal = TRUE” contributed positively to this prediction. These individualized visual explanations help interpret how personal risk profiles shape the model’s prediction at a granular level.

Figure 12. Local Interpretable Model-agnostic Explanations plot for the lung cancer model.



3.2.4 Regression tree analysis

Figure 13 presents the results of a regression tree analysis aimed at predicting lung cancer risk based on various behavioral predictors. The predictors evaluated in this analysis include BMI, the frequency of usual consumption of a portion of raw or cooked vegetables (150 g), the frequency of usual consumption of a portion of fresh fruit (150 g), the frequency of usual consumption of a portion of white meat (100 g), and the frequency of usual consumption of a portion of red meat (100 g). Additionally, cold cuts, cured meats, and sausages (50 g) consumption, alcohol consumption (e.g., glass of wine, beer, liquor), and whether the participant has had any chest diagnostic tests performed in the last year were considered. Other factors evaluated include current drug therapy, smoking status, age at which smoking started, total years smoked, pack/years, and smoking cigars or pipes. The analysis also assessed secondhand smoke exposure, including the source and hours per day, as well as whether the participant has lived in a big city or near one for more than 10 years. Occupational history with chemicals, and recent medical exams (such as Pap smear, mammography, colonoscopy, sigmoidoscopy, urological exam, PSA test, cardiological exam, and dermatological exam) were included as predictors. Additionally, we considered whether the participant had suffered from lung diseases that limited

their daily activities for more than a week in the past year, as well as the Fagerstrom test score, HADS Anxiety categories, and HADS Depression categories.

The chart provided insights into how these variables interacted with one another and their relationship to lung cancer risk. Each decision node represented a branching point, with the percentages indicating the proportion of individuals predicted to have lung cancer (TRUE) or not (FALSE) based on these factors.

The first split in the tree was based on pack-years, with the threshold set at 50.5 pack-years. Individuals who had smoked less than 50.5 pack-years were directed to the left, where 64% of participants were classified as not at risk for lung cancer. For individuals with 50.5 or more pack-years, 36% were classified as at risk for lung cancer.

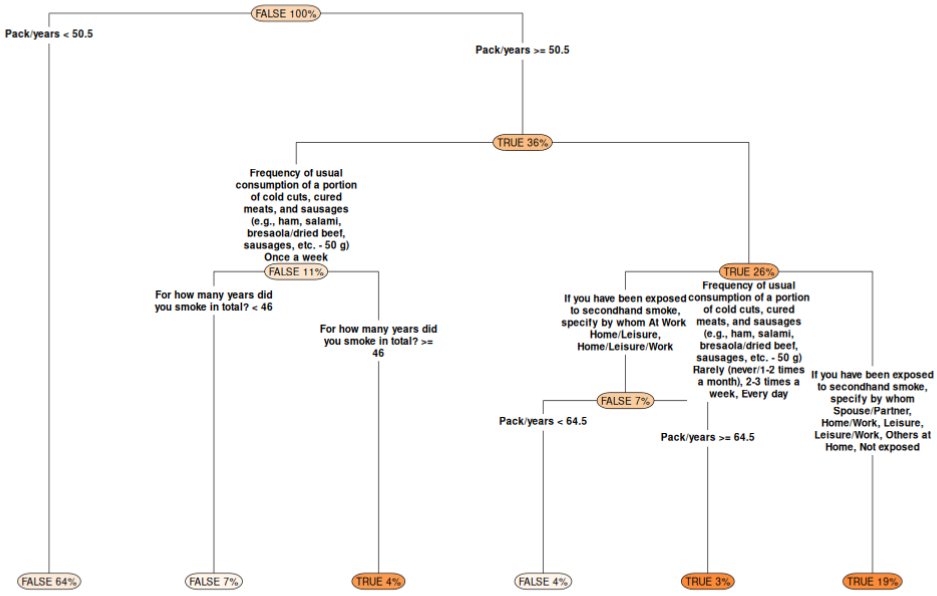
The next decision node examined the frequency of cold cut, cured meat, and sausage consumption. Those who consumed these items once a week or less were routed to the left, where only 11% were classified as not at risk for lung cancer. This indicated that, even with lower processed meat consumption, the majority of individuals in this category were still at risk. A further split occurred based on smoking duration. Individuals who had smoked for less than 46 years were routed left, where 7% were classified as not at risk for lung cancer. For individuals who had smoked for 46 years or more, 4% were classified as at risk for lung cancer, showing that a longer smoking history contributed to higher lung cancer risk.

Within the 50.5+ pack-years group, individuals with higher frequencies of cold cut, cured meat, and sausage consumption exhibited 26% at risk for lung cancer. Diving deeper into this group, 7% of individuals exposed to secondhand smoke at work, home, or during leisure activities had their lung cancer diminished. Furthermore, the regression tree was continued by splitting again based on pack-years. Those with less than 64.5 pack-years were routed left, where 4% were classified as not at risk for lung cancer. In contrast, individuals with 64.5 or more pack-years were routed to the right, where 3% were classified as at risk for lung cancer.

Finally, individuals with 50.5 or more pack-years, frequent consumption of cold cuts, cured meats, and sausages (rated as rare, everyday, or 2-3 times a week), and exposure to secondhand smoke by a spouse/partner at home, work, or leisure, or those with no exposure, were routed right, where 19% were classified as at risk for lung cancer. This suggested that high cumulative smoking exposure, secondhand smoke exposure, and frequent consumption of processed meats were associated with the second-highest risk of lung cancer.

In summary, the regression tree analysis highlighted the significant interactions between pack-years, processed meat consumption, and secondhand smoke exposure, with pack-years consistently being the most influential variable in predicting lung cancer risk. The analysis showed that while high smoking exposure and secondhand smoke exposure elevated the risk of lung cancer, a substantial proportion of individuals in these high-risk categories still did not develop lung cancer. This finding underscored the complexity of predicting lung cancer risk, where factors like smoking intensity, duration, diet, and environmental exposures interacted in shaping outcomes.

Figure 13. Results of the regression tree analysis for the COSMOS data.



3.2.5 Shiny Application for lung cancer

Figure 14 illustrates the Shiny web application developed to support end-users in quickly evaluating their lung cancer risk factors in a user-friendly interface. Upon launching the app, users are guided through a series of categorical questions that mirror the variables included in the predictive model. After completing the questionnaire, they can click the “Calculate your cancer risk factors” button to view a personalized output. This output appears as a bar chart, with each bar representing the influence of a specific factor on the overall risk estimate. An explanatory note is included to help users interpret the chart, alongside a disclaimer emphasizing that the tool is predictive and does not provide a definitive cancer diagnosis. Although we considered integrating the tool into the iBeChange platform, the current predictive accuracy of the models prompted us to assess the viability of this integration further. The app has been designed for broad accessibility and ease of use, with planned translations into Italian, Spanish, and Romanian to serve participants across multiple regions.

Figure 14. Visual Overview of the Lung Cancer Risk Prediction Application.


Risk factors for lung cancer
Please select the options below that apply to you
What is your BMI category?
Healthy Weight (18.5 to 24.9)
Overweight (25 to 29.9)
Obese (30 or greater)
How often do you usually consume a portion of raw or cooked vegetables (150g)?
Once a week
2-3 times a week
Every day
Several times a day
How often do you usually consume a portion of fresh fruit (150g)?
2-3 times a week
Every day
Several times a day
...
What is your Fagerstrom test score (nicotine dependence) category?
Very low nicotine dependence (0-2)
Low nicotine dependence (3-4)
Moderate nicotine dependence (5)
High nicotine dependence (6-7)
What is your HADS Anxiety category?
Normal
Borderline abnormal
What is your HADS Depression category?
Normal (HADS-D < 8)
Depression or borderline depression (HADS-D >= 8)
Calculate your cancer risk factors
Calculate your cancer risk factors
The plot below shows how your answers to the questionnaire above relate to your risk of developing lung cancer. The bars in the plot represent your answers, and their direction tells us whether your choices are good or bad for your health. The blue bars are positive, meaning your answers suggest behaviors that lower your risk, like your answer 'FALSE' to the question 'Smoked Cigars: Yes'. The red bars show areas where your answers suggest higher cancer risk, such as your answer 'FALSE' to the question 'Drug Therapy: No'.
Pack/Years: Above 60 = FALSE
Smoked Cigars: Yes = FALSE
Pack/Years: 2-35 = TRUE
Years Smoked: 45-61 years = FALSE
Processed Meats: Once a week = TRUE
Pack/Years: 45-60 = FALSE
Drug Therapy: No = FALSE
0.0
0.1
0.2
0.3
This is a predictive model that estimates your risk of developing lung cancer based on the answers you provided. The result is an estimate and may not always be accurate. While the model suggests behaviors that may affect your cancer risk, it doesn't guarantee that you will or won't develop cancer. The model considers factors such as smoking habits, fruit and vegetable consumption, anxiety and depression symptoms. However, real-life health outcomes are influenced by many factors that may not be captured in this model. We recommend discussing these results with a healthcare professional who can provide more personalized advice and guidance based on your individual health history and circumstances.

4. Conclusions

This study identified key sociodemographic, behavioral, clinical, psychological, and dietary factors associated with CRC risk among participants in a screening cohort.

Gender-specific risk profiles:

While male CRC cases exhibited significantly higher smoking exposure and alcohol intake, female cases were more likely to have diabetes and higher waist circumference. These gender-specific differences suggest that tailored risk-reduction strategies may be more effective than a single standard approach.

Smoking and age as risk factors:

Smoking behavior (particularly current smoking status) emerged as one of the most consistent and significant predictors of CRC, especially among male participants. Older age was also strongly associated with higher CRC risk in both genders, reaffirming established epidemiological patterns.

Dietary intake:

Diet also played an important role, with male CRC cases consuming fewer protective foods (e.g., nuts) and more alcohol, while female cases consumed less dairy. Among all dietary factors, the willingness to reduce meat intake was the most dominant predictor in tree-based models.

Behavioral readiness:

One of the most novel and actionable findings is that participants' willingness to adopt healthy behaviors (e.g., reducing meat consumption, increasing physical activity) was inversely associated with CRC risk. These motivational factors were not only independently predictive but were also ranked among the most important features in machine learning models and regression tree analysis.

Psychological and lifestyle variables:

Variables such as cancer-related worry, emotional impact, and intentions to make lifestyle changes contributed significantly to predictive accuracy. This underscores the value of including psychological and motivational dimensions in CRC risk assessment tools.

Machine learning:

The Naïve Bayes algorithm outperformed other models (e.g., LDA, GLMNet, XGBoost) in predicting CRC onset, achieving the highest F1-score and AUC. Although the predictive performance is modest and not yet clinically deployable, the models provide valuable direction for refining future CRC risk tools.

Implications for Intervention:

These results highlight the need to implement this information in the studies that will be carried out for the intervention. Therefore, the main study should prioritize:

- Smoking cessation (especially in males)
- Promoting willingness and motivation to change behavior
- Dietary modification (especially reducing meat and alcohol intake)

- Addressing comorbidities like diabetes (particularly in females)
- Engaging participants emotionally and psychologically to foster readiness for change

In addition, COSMOS analyses provided key insights into the behavioral, environmental, and psychosocial determinants of lung cancer, which are directly aligned with the objectives of Task 3.1 and the broader aim of the project: to understand how lifestyle and psychosocial risk factors contribute to cancer onset and how this knowledge can inform the development of personalized interventions.

Smoking as dominant risk factor:

Cumulative tobacco exposure was the strongest and most consistent predictor of lung cancer across all statistical and machine learning models. Participants with more than 60 pack-years had nearly fivefold increased odds of developing lung cancer (OR = 4.8), and those with over 45 years of smoking history also showed significantly higher risk. Machine learning feature importance analyses confirmed "pack-years above 60" as the top-ranked predictor.

Respiratory symptoms and clinical history:

Symptoms such as wheezing and chronic bronchitis were significantly associated with lung cancer. Wheezing episodes lasting several days or nights (OR = 2.2 to 4.5) and chronic bronchitis (OR = 2.35) demonstrated robust associations. Cough, particularly when reported as daily or intermittent, was also a relevant clinical indicator (OR = 1.98–1.91), supporting the role of respiratory history in risk stratification.

Additional predictors and environmental exposures:

Dietary and environmental factors contributed to risk prediction. Daily consumption of processed meats such as cold cuts and sausages was associated with markedly increased lung cancer risk (OR = 9.3). Secondhand smoke exposure, especially in occupational settings, and environmental or occupational history were also identified as relevant contributors. Regression tree analyses highlighted interactions among smoking exposure, processed meat intake, and secondhand smoke as synergistic factors.

Behavioral and psychological components:

Anxiety symptoms measured via the HADS scale were associated with increased lung cancer risk. Individuals categorized as "borderline abnormal" showed significantly higher odds (OR = 2.52), indicating that even mild psychological symptoms may have predictive value. These findings support the inclusion of psychosocial factors in multifactorial risk assessment models.

Machine learning performance:

Among the machine learning approaches tested, extreme gradient boosting (XGBoost) yielded the highest performance (AUC = 0.710; F1-score = 0.540). The most influential predictors included cumulative smoking exposure, processed meat consumption, and absence of ongoing drug therapy.

In sum, COSMOS analyses strengthen the evidence base for data-driven, user-centered interventions, reinforcing the Task's aim to develop personalized, predictive, and motivational tools that account for both risk behaviors and the psychological readiness to change.

5. References

1. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024; 74: 229–263.

2. Miller KD, Nogueira L, Devasia T, et al. Cancer treatment and survivorship statistics, 2022. *CA Cancer J Clin* 2022; 72: 409–436.

3. World Cancer Research Fund/American Institute for Cancer Research. *Continuous Update Project Expert Report 2018. Diet, nutrition, physical activity and colorectal cancer*, http://dietandcancerreport.org/ (2018).

4. Wooldrage K, Robbins EC, Duffy SW, et al. Long-term effects of once-only flexible sigmoidoscopy screening on colorectal cancer incidence and mortality: 21-year follow-up of the UK Flexible Sigmoidoscopy Screening randomised controlled trial. *Lancet Gastroenterol Hepatol* 2024; 9: 811–824.

5. Castells A, Quintero E, Bujanda L, et al. Effect of invitation to colonoscopy versus faecal immunochemical test screening on colorectal cancer mortality (COLONPREV): a pragmatic, randomised, controlled, non-inferiority trial. *Lancet Lond Engl* 2025; 405: 1231–1239.

6. PDQ Screening and Prevention Editorial Board. Colorectal Cancer Screening (PDQ®): Health Professional Version. In: *PDQ Cancer Information Summaries*. Bethesda (MD): National Cancer Institute (US), http://www.ncbi.nlm.nih.gov/books/NBK65825/ (2002, accessed 16 July 2025).

7. Navarro M, Nicolas A, Ferrandez A, et al. Colorectal cancer population screening programs worldwide in 2016: An update. *World J Gastroenterol* 2017; 23: 3632–3642.

8. Binefa G, Garcia M, Milà N, et al. Colorectal Cancer Screening Programme in Spain: Results of Key Performance Indicators After Five Rounds (2000-2012). *Sci Rep* 2016; 6: 19532.

9. Ji Y, Zhang Y, Liu S, et al. The epidemiological landscape of lung cancer: current status, temporal trend and future projections based on the latest estimates from GLOBOCAN 2022. *J Natl Cancer Cent* 2025; 5: 278–286.

10. Bray F, Laversanne M, Sung H, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2024; 74: 229–263.

11. Lancaster HL, Heuvelmans MA, Oudkerk M. Low-dose computed tomography lung cancer screening: Clinical evidence and implementation research. *J Intern Med* 2022; 292: 68–80.

12. Chen H-H, Wu Y-J, Wu F-Z. Precision Medicine in Lung Cancer Screening: A Paradigm Shift in Early Detection-Precision Screening for Lung Cancer. *Diagn Basel Switz* 2025; 15: 1562.

13. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Studies in

Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 2007; 147: 573–577.

14. Mansournia MA, Collins GS, Nielsen RO, et al. A CHecklist for statistical Assessment of Medical Papers (the CHAMP statement): explanation and elaboration. *Br J Sports Med*.

15. R Core Team, R. R: A Language and Environment for Statistical Computing.

16. Obón-Santacana M, Mas-Lloret J, Bars-Cortina D, et al. Meta-Analysis and Validation of a Colorectal Cancer Risk Prediction Model Using Deep Sequenced Fecal Metagenomes. *Cancers* 2022; 14: 4214.

17. Duvvuri A, Chandrasekar VT, Srinivasan S, et al. Risk of Colorectal Cancer and Cancer Related Mortality After Detection of Low-risk or High-risk Adenomas, Compared With No Adenoma, at Index Colonoscopy: A Systematic Review and Meta-analysis. *Gastroenterology* 2021; 160: 1986-1996.e3.

18. Roshandel G, Ghasemi-Kebria F, Malekzadeh R. Colorectal Cancer: Epidemiology, Risk Factors, and Prevention. *Cancers* 2024; 16: 1530.

19. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983; 67: 361–370.

20. Kuhn M, Johnson K, others. *Applied predictive modeling*. Springer, 2013.

21. Prantner B. Visualization of imputed values using the R-package VIM.

22. Buuren S van, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *J Stat Softw* 2011; 45: 1–67.

23. Van Buuren S, Brand JP, Groothuis-Oudshoorn CG, et al. Fully conditional specification in multivariate imputation. *J Stat Comput Simul* 2006; 76: 1049–1064.

24. Chawla NV, Bowyer KW, Hall LO, et al. SMOTE: Synthetic Minority Over-sampling Technique. *J Artif Intell Res* 2002; 16: 321–357.

25. Torgo L. *Data Mining with R: Learning with Case Studies*. New York: Chapman and Hall/CRC, 2011. Epub ahead of print 22 September 2011. DOI: 10.1201/9780429292859.

26. Friedman J, Hastie T, Tibshirani R. *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001.

27. Ribeiro MT, Singh S, Guestrin C. ‘Why Should I Trust You?’: Explaining the Predictions of Any Classifier. Epub ahead of print 9 August 2016. DOI: 10.48550/arXiv.1602.04938.

28. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2019.

29. White MC, Holman DM, Boehm JE, et al. Age and cancer risk: a potentially modifiable relationship. *Am J Prev Med* 2014; 46: S7–S15.

30. Aune D, Giovannucci E, Boffetta P, et al. Fruit and vegetable intake and the risk of cardiovascular disease, total cancer and all-cause mortality-a systematic review and dose-response meta-analysis of prospective studies. *Int J Epidemiol* 2017; 46: 1029–1056.

31. Aune D, Keum N, Giovannucci E, et al. Nut consumption and risk of cardiovascular disease, total cancer, all-cause and cause-specific mortality: a systematic review and dose-response meta-analysis of prospective studies. *BMC Med* 2016; 14: 207.

32. Zigmond AS, Snaith RP. The hospital anxiety and depression scale. *Acta Psychiatr Scand* 1983; 67: 361–370.

6. Appendices (Tables)

Supplemental Table 1. CRC sample description, including additional detail on numeric variables.

	FEMALES				MALES			
Variables	**Controls (n=430)**	**Cases (n=76)**	**p-value**	**Missing (%)**	**Controls (n=420)**	**Cases (n=148)**	**p-value**	**Missing (%)**
n	430	76			420	148		
Age at recruitment (mean (SD))	59.65 (5.74)	61.55 (5.33)	0.007	0.0	60.00 (5.74)	61.53 (5.71)	0.005	0.0
Age at recruitment (median [IQR])	60.00 [55.00, 65.00]	63.00 [57.75, 65.25]	0.009	0.0	60.00 [55.00, 65.00]	63.00 [57.00, 67.00]	0.005	0.0
Age at recruitment (%)			0.065	0.0			0.037	0.0
- 49 - 55	126 (29.3)	13 (17.1)			108 (25.7)	25 (16.9)		
- 56 - 60	101 (23.5)	15 (19.7)			115 (27.4)	34 (23.0)		
- 61 - 65	121 (28.1)	29 (38.2)			103 (24.5)	44 (29.7)		
- 66 - 70	82 (19.1)	19 (25.0)			94 (22.4)	45 (30.4)		
Ethnicity or race group: White/Caucasian (%)	418 (97.2)	74 (97.4)	1.000	0.0	411 (97.9)	148 (100.0)	0.158	0.0
Education level (%)			0.077	0.0			0.144	0.0
- University	54 (12.6)	2 (2.6)			73 (17.4)	28 (18.9)		
- High school diploma (BUP or COU)	68 (15.8)	13 (17.1)			69 (16.4)	31 (20.9)		
- Vocational training (FP, or similar)	82 (19.1)	19 (25.0)			100 (23.8)	26 (17.6)		

- Complete primary education (EGB, or similar)	189 (44.0)	31 (40.8)			167 (39.8)	54 (36.5)		
- Incomplete primary education	30 (7.0)	8 (10.5)			10 (2.4)	7 (4.7)		
- No formal education, but can read	7 (1.6)	3 (3.9)			1 (0.2)	2 (1.4)		
Marital status (%)			0.180	0.2			0.387	0.0
- Single/never married	27 (6.3)	4 (5.3)			26 (6.2)	5 (3.4)		
- Married or living with a partner	329 (76.7)	52 (68.4)			345 (82.1)	120 (81.1)		
- Separated or divorced	45 (10.5)	10 (13.2)			42 (10.0)	19 (12.8)		
- Widowed	28 (6.5)	10 (13.2)			7 (1.7)	4 (2.7)		
Social class of parents (%)			0.425	0.6			0.996	0.2
- Upper social class	5 (1.2)	0 (0.0)			6 (1.4)	2 (1.4)		
- Middle social class	240 (56.2)	39 (51.3)			228 (54.4)	81 (54.7)		
- Lower social class	182 (42.6)	37 (48.7)			185 (44.2)	65 (43.9)		
Number of siblings (mean (SD))	3.23 (2.54)	3.00 (2.47)	0.476	0.2	2.96 (2.40)	2.55 (1.77)	0.055	0.0
Number of siblings (median [IQR])	3.00 [1.00, 4.75]	2.00 [1.00, 4.00]	0.414	0.2	2.00 [1.00, 4.00]	2.00 [1.00, 4.00]	0.239	0.0

Number of children (mean (SD))	1.90 (1.09)	1.81 (1.39)	0.523	0.2	1.77 (0.89)	1.84 (1.04)	0.411	0.0
Number of children (median [IQR])	2.00 [1.00, 2.00]	2.00 [1.00, 2.00]	0.135	0.2	2.00 [1.00, 2.00]	2.00 [1.00, 2.00]	0.601	0.0
Weight (mean (SD))	69.75 (13.51)	70.59 (12.39)	0.612	0.2	84.67 (13.43)	83.81 (14.50)	0.511	0.0
Weight (median [IQR])	68.00 [60.00, 77.00]	70.00 [62.00, 77.25]	0.348	0.2	83.00 [75.00, 92.25]	82.50 [74.00, 93.00]	0.485	0.0
Occupation (%)			0.269	0.0			0.472	0.2
- Working	179 (41.6)	26 (34.2)			203 (48.3)	63 (42.9)		
- Unemployed	61 (14.2)	10 (13.2)			27 (6.4)	12 (8.2)		
- Housewife or domestic worker	78 (18.1)	21 (27.6)						
- Retired	112 (26.0)	19 (25.0)			190 (45.2)	72 (49.0)		
Physical activity at work (%)			0.192	2.8			0.092	0.2
- Sedentary	65 (15.4)	11 (15.7)			61 (14.5)	17 (11.6)		
- Slightly active	80 (19.0)	6 (8.6)			65 (15.5)	18 (12.2)		
- Moderately active	93 (22.0)	14 (20.0)			109 (26.0)	50 (34.0)		
- Fairly active	129 (30.6)	29 (41.4)			130 (31.0)	35 (23.8)		
- Very active	55 (13.0)	10 (14.3)			55 (13.1)	27 (18.4)		
METs hours per week (%)			0.015	0.0			0.259	1.2
- 0	141 (32.8)	24 (31.6)			94 (22.8)	45 (30.4)		
- 0.01 - 17.4	112 (26.0)	11 (14.5)			82 (19.9)	30 (20.3)		

- 17.41 - 31	111 (25.8)	19 (25.0)			104 (25.2)	30 (20.3)		
- 31.01 - 140	66 (15.3)	22 (28.9)			133 (32.2)	43 (29.1)		
METs hours per week walking (%)			0.044	0.0			0.843	0.2
- 0	238 (55.3)	42 (55.3)			222 (53.0)	75 (50.7)		
- 0.01 - 18	101 (23.5)	10 (13.2)			88 (21.0)	31 (20.9)		
- 18.01 - 108	91 (21.2)	24 (31.6)			109 (26.0)	42 (28.4)		
Waist circumference (mean (SD))	91.47 (13.00)	94.94 (11.40)	0.037	9.5	99.29 (10.92)	100.43 (12.31)	0.323	10.6
Waist circumference (median [IQR])	91.00 [83.00, 100.00]	94.50 [88.00, 103.00]	0.024	9.5	98.00 [91.00, 106.00]	101.00 [93.00, 109.00]	0.153	10.6
Waist circumference (%)			0.203	9.5			0.312	10.6
- 0 - 88	166 (42.8)	23 (32.9)			54 (14.2)	19 (14.7)		
- 88.01 - 96	91 (23.5)	15 (21.4)			99 (26.1)	24 (18.6)		
- 96.01 - 104	68 (17.5)	19 (27.1)			110 (29.0)	38 (29.5)		
- 104.01 - 137	63 (16.2)	13 (18.6)			116 (30.6)	48 (37.2)		
Hip circumference (mean (SD))	103.66 (12.01)	105.49 (10.19)	0.240	12.1	103.10 (8.22)	103.16 (11.36)	0.946	15.3
Hip circumference (median [IQR])	103.00 [96.00, 109.00]	105.00 [99.50, 112.00]	0.083	12.1	103.00 [98.00, 108.00]	102.00 [98.00, 108.00]	0.566	15.3
Hip circumference (%)			0.079	12.1			0.490	15.3

- 0 - 98	134 (35.5)	17 (25.0)			97 (27.1)	39 (31.7)		
- 98.01 - 103	62 (16.4)	7 (10.3)			91 (25.4)	30 (24.4)		
- 103.01 - 109	87 (23.1)	23 (33.8)			98 (27.4)	26 (21.1)		
- 109.01 - 177	94 (24.9)	21 (30.9)			72 (20.1)	28 (22.8)		
Waist-hip ratio (mean (SD))	0.88 (0.08)	0.90 (0.08)	0.077	12.1	0.97 (0.07)	0.97 (0.08)	0.327	15.3
Waist-hip ratio (median [IQR])	0.89 [0.83, 0.93]	0.89 [0.84, 0.95]	0.140	12.1	0.96 [0.92, 1.02]	0.98 [0.94, 1.03]	0.088	15.3
Waist-hip ratio (%)	18 (4.8)	8 (11.8)	0.048	12.1	115 (32.1)	52 (42.3)	0.054	15.3
Weight 1 year ago (mean (SD))	69.91 (14.60)	70.65 (12.95)	0.681	1.8	84.59 (13.57)	84.07 (14.65)	0.700	0.7
Weight 1 year ago (median [IQR])	67.00 [60.00, 76.00]	70.00 [60.00, 78.50]	0.374	1.8	83.00 [75.00, 93.25]	82.00 [73.75, 93.00]	0.583	0.7
Maximum weight (mean (SD))	74.77 (16.36)	75.56 (13.62)	0.695	1.0	89.73 (15.20)	89.14 (15.19)	0.686	0.7
Maximum weight (median [IQR])	71.00 [64.00, 82.00]	74.00 [65.50, 83.50]	0.261	1.0	88.00 [79.00, 97.00]	86.50 [80.00, 98.00]	0.579	0.7
Age at maximum weight (mean (SD))	55.57 (13.71)	55.13 (13.85)	0.799	1.4	55.06 (12.58)	56.53 (15.21)	0.250	0.5
Age at maximum weight (median [IQR])	56.00 [50.00, 63.00]	57.00 [50.00, 64.00]	0.306	1.4	55.00 [50.00, 62.00]	59.00 [50.00, 64.75]	0.092	0.5
Current height (cm) (mean (SD))	157.59 (6.13)	157.61 (5.51)	0.979	0.2	171.12 (6.71)	170.82 (6.84)	0.635	0.0

Current height (cm) (median [IQR])	157.00 [153.00, 162.00]	158.00 [153.75, 160.00]	0.992	0.2	171.00 [166.00, 175.00]	170.00 [167.00, 175.00]	0.402	0.0
Current height (cm) (%)			0.274	0.0			0.129	0.0
- 150 or less	58 (13.5)	8 (10.5)						
- 151 - 160	243 (56.5)	52 (68.4)			22 (5.2)	8 (5.4)		
- 161 - 170	118 (27.4)	15 (19.7)			171 (40.7)	74 (50.0)		
- 171 or more	11 (2.6)	1 (1.3)			227 (54.0)	66 (44.6)		
BMI (mean (SD))	28.09 (5.24)	28.46 (4.97)	0.566	0.4	28.89 (4.07)	28.67 (4.40)	0.586	0.0
BMI (median [IQR])	27.24 [24.44, 30.85]	28.25 [25.54, 30.95]	0.288	0.4	28.37 [26.09, 31.60]	27.89 [25.93, 30.86]	0.527	0.0
BMI (%)			0.568	0.4			0.401	0.0
- Underweight (< 18.5)	2 (0.5)	0 (0.0)						
- Normal weight (18.5 - 24.9)	131 (30.6)	18 (23.7)			69 (16.4)	31 (20.9)		
- Overweight (25 - 29.9)	166 (38.8)	34 (44.7)			208 (49.5)	66 (44.6)		
- Obesity (>= 30)	129 (30.1)	24 (31.6)			143 (34.0)	51 (34.5)		
In your lifetime, have you ever smoked? 'YES' means at least 100 cigarettes or 360 grams of tobacco in your lifetime. (%)	204 (47.4)	38 (50.0)	0.774	0.0	320 (76.2)	130 (87.8)	0.004	0.0

Have you ever smoked regularly, i.e., at least one cigarette per day for six months or more? (%)	203 (47.2)	38 (50.0)	0.746	0.0	320 (76.2)	130 (87.8)	0.004	0.0
Age at smoking initiation (mean (SD))	18.00 (4.60)	20.29 (7.33)	0.012	0	17.20 (4.13)	16.79 (3.08)	0.306	0.2
Age at smoking initiation (median [IQR])	17.00 [15.00, 19.00]	18.00 [15.25, 23.00]	0.082	0	17.00 [15.00, 18.25]	17.00 [15.00, 18.00]	0.691	0.2
Age at smoking initiation (%)			0.139	0.0			0.047	0.2
- 8 - 15	53 (12.3)	10 (13.2)			110 (26.2)	44 (29.9)		
- 15 - 17	52 (12.1)	6 (7.9)			76 (18.1)	35 (23.8)		
- 17 - 19	50 (11.6)	6 (7.9)			68 (16.2)	24 (16.3)		
- 19 - 54	48 (11.2)	16 (21.1)			66 (15.7)	26 (17.7)		
- Never smoked	227 (52.8)	38 (50.0)			100 (23.8)	18 (12.2)		
Current smoker (%)	87 (20.3)	16 (21.1)	1.000	0.2	91 (21.7)	66 (44.6)	<0.001	0.0
Number of cigarettes on average, excluding non-smokers (mean (SD))	14.55 (10.19)	15.25 (4.80)	0.789	27.3	14.90 (9.97)	13.12 (10.64)	0.285	51.6
Number of cigarettes on average, excluding non-smokers (median [IQR])	12.00 [9.00, 20.00]	15.00 [13.75, 20.00]	0.315	27.3	15.00 [7.50, 20.00]	10.00 [5.00, 20.00]	0.171	51.6

Number of cigarettes on average (%)			0.208	27.3			<0.001	51.6
- Never smoked	227 (72.3)	38 (70.4)			100 (52.4)	18 (21.4)		
- 1 - 8	22 (7.0)	1 (1.9)			24 (12.6)	27 (32.1)		
- 9 - 15	30 (9.6)	9 (16.7)			26 (13.6)	16 (19.0)		
- 16 - 20	27 (8.6)	6 (11.1)			29 (15.2)	13 (15.5)		
- 21 - 60	8 (2.5)	0 (0.0)			12 (6.3)	10 (11.9)		
Current frequency of smoking (%)			0.913	0.2			<0.001	0.0
- Day	83 (19.3)	16 (21.1)			88 (21.0)	64 (43.2)		
- Week	3 (0.7)	0 (0.0)			2 (0.5)	2 (1.4)		
- Month	1 (0.2)	0 (0.0)			1 (0.2)	0 (0.0)		
- Former smoker	116 (27.0)	22 (28.9)			229 (54.5)	64 (43.2)		
- Never	226 (52.7)	38 (50.0)			100 (23.8)	18 (12.2)		
Passive smoker (%)	114 (31.1)	21 (32.8)	0.905	15.0	88 (26.4)	37 (32.5)	0.264	21.3
Smoking status (%)			0.901	0.0			<0.001	0.2
- Never	227 (52.8)	38 (50.0)			99 (23.6)	18 (12.2)		
- Ex-Smoker	116 (27.0)	22 (28.9)			229 (54.7)	64 (43.2)		
- Smoker	87 (20.2)	16 (21.1)			91 (21.7)	66 (44.6)		
Pack years, excluding never-smokers (mean (SD))	23.48 (21.48)	25.81 (17.26)	0.534	1.2	32.26 (25.39)	38.77 (30.19)	0.026	5.6

Pack years, excluding never-smokers (median [IQR])	19.31 [5.03, 34.02]	25.52 [15.71, 39.03]	0.212	1.2	25.82 [11.41, 45.03]	33.02 [17.01, 50.03]	0.030	5.6
Pack years (%)			0.502	1.2			0.007	5.6
- Never smoked	227 (53.4)	38 (50.7)			100 (24.9)	18 (13.3)		
- 0.09 - 11.01	68 (16.0)	8 (10.7)			73 (18.2)	17 (12.6)		
- 11.02 - 25.22	48 (11.3)	10 (13.3)			75 (18.7)	28 (20.7)		
- 25.23 - 42.03	53 (12.5)	14 (18.7)			64 (16.0)	33 (24.4)		
- 42.04 - 168.12	29 (6.8)	5 (6.7)			89 (22.2)	39 (28.9)		
Average lifetime intensity in cigarettes/year, excluding never-smokers (mean (SD))	5515.59 (4184.70)	5662.66 (3392.65)	0.838	0.2	55682.16 (659114.94)	8053.66 (5941.32)	0.431	4.9
Average lifetime intensity in cigarettes/year, excluding never-smokers (median [IQR])	4383.00 [2192.00 , 7305.00]	5478.00 [3652.00 , 7305.00]	0.450	0.2	7305.00 [3652.00, 10958.00]	7305.00 [3652.00, 10945.50]	0.381	4.9
Average lifetime intensity in cigarettes/year (%)			0.733	0.2			0.010	4.9
- Never smoked	227 (52.9)	38 (50.0)			100 (24.8)	18 (13.1)		
- 36 - 3652	98 (22.8)	15 (19.7)			85 (21.1)	32 (23.4)		
- 3653 - 7305	74 (17.2)	18 (23.7)			108 (26.8)	53 (38.7)		
- 7306 - 9131	4 (0.9)	1 (1.3)			11 (2.7)	1 (0.7)		
- 9132 - 10957625	26 (6.1)	4 (5.3)			99 (24.6)	33 (24.1)		

Years of smoking, excluding non-smokers (mean (SD))	29.31 (12.86)	30.43 (12.05)	0.623	1.2	29.30 (12.25)	36.41 (10.65)	<0.001	2.1
Years of smoking, excluding non-smokers (median [IQR])	33.00 [20.00, 39.00]	34.00 [23.00, 39.00]	0.643	1.2	31.00 [20.00, 38.75]	38.50 [30.00, 45.00]	<0.001	2.1
Years of smoking (%)			0.732	1.2			<0.001	2.1
- Never smoked	227 (53.4)	38 (50.7)			100 (24.2)	18 (12.7)		
- 1 - 22	55 (12.9)	8 (10.7)			97 (23.4)	15 (10.6)		
- 22.01 - 33	46 (10.8)	10 (13.3)			90 (21.7)	28 (19.7)		
- 33.01 - 40	54 (12.7)	13 (17.3)			65 (15.7)	30 (21.1)		
- 40.01 - 56	43 (10.1)	6 (8.0)			62 (15.0)	51 (35.9)		
Heartburn (%)	182 (42.4)	27 (35.5)	0.318	0.2	136 (32.5)	42 (28.8)	0.469	0.5
Medication for heartburn (%)	179 (41.6)	27 (35.5)	0.384	0.0	155 (36.9)	45 (30.4)	0.186	0.0
Laxative use (%)	121 (28.5)	21 (28.4)	1.000	1.6	44 (10.6)	8 (5.5)	0.098	0.9
Diabetes = Yes (%)	38 (8.8)	18 (23.7)	<0.001	0.0	63 (15.0)	20 (13.5)	0.760	0.0
Hypertension = Yes (%)	136 (31.6)	26 (34.2)	0.755	0.0	180 (42.9)	72 (48.6)	0.261	0.0
High cholesterol = Yes (%)	136 (31.6)	32 (42.1)	0.098	0.0	167 (39.9)	65 (43.9)	0.443	0.2
Angina pectoris (%)	7 (1.6)	0 (0.0)	0.557	0.0	14 (3.3)	5 (3.4)	1.000	0.0
Myocardial infarction (%)	5 (1.2)	0 (0.0)	0.752	0.0	14 (3.3)	9 (6.1)	0.224	0.0
Stroke (%)	11 (2.6)	2 (2.6)	1.000	0.0	16 (3.8)	6 (4.1)	1.000	0.0

Circulatory problems (%)	53 (12.3)	6 (7.9)	0.360	0.0	38 (9.1)	14 (9.5)	1.000	0.4
Arthritis (%)	128 (29.8)	28 (36.8)	0.279	0.2	79 (18.9)	29 (19.6)	0.940	0.2
Migraine (%)	68 (15.8)	10 (13.2)	0.675	0.0	23 (5.5)	9 (6.1)	0.951	0.2
Anemia (%)	55 (12.8)	8 (10.5)	0.717	0.0	11 (2.6)	3 (2.0)	0.933	0.4
Diverticulitis (%)	5 (1.2)	2 (2.7)	0.622	0.2	8 (1.9)	2 (1.4)	0.944	0.4
Celiac disease (%)	5 (1.2)	0 (0.0)	0.750	0.4	0 (0.0)	0 (0.0)		
Depression (%)	126 (29.4)	26 (34.7)	0.439	0.6	53 (12.6)	23 (15.6)	0.438	0.4
Osteoporosis (%)	51 (11.9)	11 (14.5)	0.652	0.0	1 (0.2)	0 (0.0)	1.000	0.2
Polyps (%)	17 (4.0)	4 (5.3)	0.829	0.0	31 (7.4)	6 (4.1)	0.224	0.0
Dyspepsia (%)	27 (6.3)	6 (7.9)	0.792	0.4	15 (3.6)	1 (0.7)	0.123	0.0
Schizophrenia (%)	1 (0.2)	0 (0.0)	1.000	0.6	2 (0.5)	0 (0.0)	0.973	0.0
Anti-inflammatory medication (%)	111 (27.4)	21 (29.2)	0.869	5.7	81 (20.7)	30 (21.3)	0.974	6.2
Menstruation status = still has periods (%)	26 (6.1)	1 (1.3)	0.161	0.6				
Age at last menstruation (mean (SD))	48.87 (4.64)	48.40 (5.09)	0.461	10.3				
Age at last menstruation (median [IQR])	50.00 [47.00, 52.00]	50.00 [45.00, 52.00]	0.387	10.3				
Age at last menstruation (%)			0.296	10.3				
- 33 - 46	90 (23.1)	21 (32.8)						

- 46 - 50	144 (36.9)	23 (35.9)						
- 50 - 52	61 (15.6)	8 (12.5)						
- 52 - 60	69 (17.7)	11 (17.2)						
- Still has periods	26 (6.7)	1 (1.6)						
Age at first menstruation (%)			0.114	1.6				
- 8 - 11	111 (26.2)	27 (36.0)						
- 11 - 13	172 (40.7)	33 (44.0)						
- 13 - 14	99 (23.4)	10 (13.3)						
- 14 - 18	41 (9.7)	5 (6.7)						
Use contraceptive (%)	294 (69.2)	47 (62.7)	0.326	1.2				
Menopause treatment (%)	50 (11.9)	11 (15.5)	0.514	3.0				
Prostate disease (%)					81 (19.4)	26 (17.7)	0.743	0.5
Weight loss (%)	6 (1.4)	1 (1.3)	1.000	0.0	4 (1.0)	0 (0.0)	0.535	0.0
Are early detection programs useful? (%)			0.181	35.0			0.501	32.7
- Strongly disagree	2 (0.7)	0 (0.0)			5 (1.8)	2 (2.0)		
- Disagree					0 (0.0)	1 (1.0)		
- Neither agree nor disagree					2 (0.7)	2 (2.0)		
- Agree	21 (7.3)	0 (0.0)			18 (6.4)	6 (5.9)		

- Strongly agree	266 (92.0)	40 (100.0)			255 (90.7)	90 (89.1)		
- NS/NC (Not sure/No comment)					1 (0.4)	0 (0.0)		
Willing to participate again? (%)	277 (99.6)	35 (97.2)	0.547	37.9	269 (99.3)	95 (99.0)	1.000	35.4
During the past month, how often have you thought about your chances of getting cancer? (%)			0.717	34.6			0.222	32.2
- Rarely or never	98 (33.9)	11 (26.2)			127 (44.9)	34 (33.3)		
- Sometimes	138 (47.8)	23 (54.8)			132 (46.6)	56 (54.9)		
- Often	42 (14.5)	7 (16.7)			19 (6.7)	9 (8.8)		
- Almost all the time	11 (3.8)	1 (2.4)			5 (1.8)	3 (2.9)		
During the past month, has thinking about the possibility of developing cancer affected your mood? (%)			0.701	34.6			0.400	32.4
- Rarely or never	155 (53.6)	21 (50.0)			178 (62.9)	56 (55.4)		
- Sometimes	104 (36.0)	17 (40.5)			91 (32.2)	41 (40.6)		
- Often	23 (8.0)	4 (9.5)			12 (4.2)	4 (4.0)		
- Almost all the time	7 (2.4)	0 (0.0)			2 (0.7)	0 (0.0)		

During the past month, has thinking about the possibility of developing cancer affected your ability to carry out your daily activities? (%)			0.785	34.6			0.706	32.6
- Rarely or never	173 (59.9)	26 (61.9)			198 (70.2)	67 (66.3)		
- Sometimes	96 (33.2)	13 (31.0)			75 (26.6)	31 (30.7)		
- Often	15 (5.2)	3 (7.1)			7 (2.5)	3 (3.0)		
- Almost all the time	5 (1.7)	0 (0.0)			2 (0.7)	0 (0.0)		
To what extent do you worry about the possibility of developing cancer one day? (%)			0.449	34.6			0.587	32.4
- Not at all	70 (24.2)	9 (21.4)			82 (29.1)	35 (34.3)		
- A little	118 (40.8)	13 (31.0)			130 (46.1)	40 (39.2)		
- Quite a bit	72 (24.9)	14 (33.3)			51 (18.1)	18 (17.6)		
- A great deal	29 (10.0)	6 (14.3)			19 (6.7)	9 (8.8)		
How often do you worry about the possibility of developing cancer? (%)			0.809	34.6			0.550	32.4
- Never or rarely	102 (35.3)	17 (40.5)			120 (42.6)	44 (43.1)		
- Occasionally	156 (54.0)	20 (47.6)			147 (52.1)	49 (48.0)		

- Frequently	29 (10.0)	5 (11.9)			13 (4.6)	7 (6.9)		
- Constantly	2 (0.7)	0 (0.0)			2 (0.7)	2 (2.0)		
Is being worried about developing cancer an important issue for you? (%)			0.989	34.6			0.895	32.4
- No; not at all	92 (31.8)	13 (31.0)			115 (40.8)	46 (45.1)		
- A little	81 (28.0)	12 (28.6)			65 (23.0)	21 (20.6)		
- Yes; it's definitely a problem	73 (25.3)	10 (23.8)			67 (23.8)	23 (22.5)		
- Yes; it's a very serious problem	43 (14.9)	7 (16.7)			35 (12.4)	12 (11.8)		
Willing to change the lifestyle to reduce colon cancer risk (%)	271 (97.1)	40 (100.0)	0.586	37.0	264 (97.8)	91 (91.9)	0.021	35.0
If you were obese, would you lose weight? (%)			0.733	35.2			0.239	34.2
- Yes	216 (75.0)	31 (77.5)			212 (77.4)	72 (72.0)		
- No	5 (1.7)	0 (0.0)			3 (1.1)	4 (4.0)		
- I'm not obese	63 (21.9)	9 (22.5)			58 (21.2)	24 (24.0)		
- Not sure	4 (1.4)	0 (0.0)			1 (0.4)	0 (0.0)		
If you were a smoker, would you quit smoking? (%)			0.508	35.4			<0.001	35.2
- Yes	86 (30.0)	16 (40.0)			97 (35.8)	52 (53.6)		
- No	12 (4.2)	1 (2.5)			6 (2.2)	9 (9.3)		

- I'm not a smoker	184 (64.1)	23 (57.5)			165 (60.9)	34 (35.1)		
- Not sure	5 (1.7)	0 (0.0)			3 (1.1)	2 (2.1)		
If you were a heavy drinker, would you reduce your alcohol consumption? (%)			0.759	35.2			0.005	34.3
- Yes	81 (28.1)	11 (27.5)			125 (45.8)	62 (62.0)		
- No	2 (0.7)	0 (0.0)			5 (1.8)	5 (5.0)		
- I drink less alcohol	199 (69.1)	29 (72.5)			138 (50.5)	31 (31.0)		
- Not sure	6 (2.1)	0 (0.0)			5 (1.8)	2 (2.0)		
If you did little exercise: would you do more exercise on a regular basis? (%)			0.535	35.2			0.026	34.2
- Yes	259 (89.9)	38 (95.0)			258 (94.2)	88 (88.0)		
- No	7 (2.4)	0 (0.0)			6 (2.2)	8 (8.0)		
- I exercise a lot	14 (4.9)	2 (5.0)			6 (2.2)	4 (4.0)		
- Not sure	8 (2.8)	0 (0.0)			4 (1.5)	0 (0.0)		
If you were to eat a meat-heavy diet: would you eat less meat? (%)			0.407	35.2			0.042	34.2
- Yes	237 (82.3)	30 (75.0)			252 (92.0)	86 (86.0)		
- No	2 (0.7)	0 (0.0)			2 (0.7)	5 (5.0)		
- I don't eat much meat	45 (15.6)	10 (25.0)			19 (6.9)	9 (9.0)		

- Not sure	4 (1.4)	0 (0.0)			1 (0.4)	0 (0.0)		
If you were to eat a diet low in vegetables: would you eat more vegetables? (%)			0.576	35.2			0.292	34.2
- Yes	243 (84.4)	32 (80.0)			247 (90.1)	91 (91.0)		
- No	4 (1.4)	1 (2.5)			6 (2.2)	5 (5.0)		
- I eat a lot of vegetables	35 (12.2)	7 (17.5)			19 (6.9)	4 (4.0)		
- Not sure	6 (2.1)	0 (0.0)			2 (0.7)	0 (0.0)		
Total energy (kcal/day) (mean (SD))	1635.82 (536.88)	1605.64 (629.26)	0.677	8.7	1983.44 (649.81)	1974.27 (572.00)	0.885	8.1
Total energy (kcal/day) (median [IQR])	1576.80 [1277.66, 1940.25]	1473.67 [1316.56, 1692.40]	0.266	8.7	1906.89 [1499.86, 2361.09]	1952.05 [1560.55, 2331.48]	0.912	8.1
Total protein (g/day) (mean (SD))	72.91 (23.48)	72.80 (26.58)	0.973	8.7	87.09 (26.64)	86.12 (25.27)	0.712	8.1
Total protein (g/day) (median [IQR])	70.73 [56.19, 84.43]	68.97 [60.86, 80.61]	0.610	8.7	84.16 [67.73, 101.33]	83.22 [70.59, 98.80]	0.810	8.1
Total carbohydrates (g/day) (mean (SD))	161.08 (62.35)	156.62 (62.66)	0.586	8.7	192.99 (71.92)	186.79 (63.95)	0.378	8.1
Total carbohydrates (g/day) (median [IQR])	150.87 [117.97, 192.87]	143.69 [125.45, 171.00]	0.458	8.7	184.56 [143.75, 229.36]	176.46 [136.40, 229.30]	0.423	8.1
Total fats (g/day) (mean (SD))	73.02 (27.94)	71.18 (33.44)	0.627	8.7	82.74 (33.46)	81.07 (29.10)	0.606	8.1

Total fats (g/day) (median [IQR])	69.58 [52.06, 90.49]	67.49 [52.25, 79.46]	0.320	8.7	76.43 [59.29, 99.45]	80.84 [58.41, 96.76]	0.963	8.1
Total fiber (g/day) (mean (SD))	20.02 (9.89)	20.62 (9.63)	0.641	8.7	19.74 (8.84)	18.30 (6.70)	0.085	8.1
Total fiber (g/day) (median [IQR])	17.86 [13.89, 24.04]	18.47 [14.50, 24.59]	0.442	8.7	17.85 [13.95, 23.34]	17.47 [13.93, 22.03]	0.250	8.1
Total ethanol (g/day) (mean (SD))	5.37 (9.37)	5.96 (9.11)	0.632	8.7	16.20 (18.58)	21.21 (22.18)	0.011	8.1
Total ethanol (g/day) (median [IQR])	1.53 [0.00, 6.70]	2.20 [0.00, 7.90]	0.608	8.7	10.30 [2.85, 22.51]	14.60 [4.92, 30.32]	0.007	8.1
Red meat (g/day) (mean (SD))	19.36 (15.18)	20.52 (16.60)	0.569	8.7	35.03 (27.76)	35.20 (30.50)	0.951	8.1
Red meat (g/day) (median [IQR])	16.78 [7.13, 27.65]	15.35 [7.69, 29.16]	0.743	8.7	29.03 [17.53, 45.64]	29.56 [16.53, 42.83]	0.677	8.1
White meat (g/day) (mean (SD))	25.91 (21.11)	23.99 (16.68)	0.476	8.7	30.46 (20.71)	30.35 (22.45)	0.959	8.1
White meat (g/day) (median [IQR])	18.76 [13.31, 35.80]	19.02 [13.45, 33.72]	0.797	8.7	23.45 [18.16, 42.09]	24.05 [17.59, 40.28]	0.881	8.1
Cured and processed meat (g/day) (mean (SD))	29.11 (20.83)	28.13 (17.99)	0.714	8.7	48.36 (31.10)	49.68 (30.25)	0.671	8.1
Cured and processed meat (g/day) (median [IQR])	24.80 [14.35, 38.06]	27.57 [15.28, 38.69]	0.937	8.7	42.15 [27.19, 61.62]	42.11 [28.87, 63.54]	0.617	8.1
All meat (g/day) (mean (SD))	75.18 (39.40)	73.45 (37.81)	0.736	8.7	115.00 (55.20)	117.29 (59.52)	0.686	8.1

All meat (g/day) (median [IQR])	72.36 [48.81, 96.01]	72.74 [47.81, 91.17]	0.746	8.7	105.40 [78.22, 141.40]	101.80 [78.31, 141.26]	0.984	8.1
White fish (g/day) (mean (SD))	16.37 (14.62)	18.32 (12.88)	0.304	8.7	14.50 (10.78)	15.76 (14.30)	0.286	8.1
White fish (g/day) (median [IQR])	14.44 [6.04, 18.74]	15.76 [9.63, 23.39]	0.073	8.7	15.73 [6.39, 19.11]	15.35 [6.29, 20.88]	0.981	8.1
Blue fish (g/day) (mean (SD))	13.82 (13.27)	15.38 (12.03)	0.367	8.7	14.71 (12.45)	14.53 (12.10)	0.884	8.1
Blue fish (g/day) (median [IQR])	13.63 [3.29, 18.08]	15.35 [6.33, 20.44]	0.134	8.7	15.43 [6.04, 19.84]	15.42 [3.37, 20.66]	0.877	8.1
Fruits (g/day) (mean (SD))	264.69 (190.40)	269.21 (166.41)	0.854	9.7	222.37 (169.01)	193.96 (138.98)	0.081	8.5
Fruits (g/day) (median [IQR])	225.35 [133.33, 359.24]	253.78 [151.70, 367.41]	0.488	9.7	183.79 [100.21, 303.53]	168.55 [85.57, 258.18]	0.177	8.5
Vegetables (g/day) (mean (SD))	195.03 (118.09)	227.75 (142.49)	0.042	9.1	150.61 (93.75)	144.04 (88.53)	0.480	8.1
Vegetables (g/day) (median [IQR])	169.14 [119.16, 245.22]	192.45 [135.93, 266.87]	0.061	9.1	130.45 [87.66, 198.28]	127.51 [82.83, 203.30]	0.544	8.1
Legumes (g/day) (mean (SD))	38.11 (34.86)	39.33 (32.06)	0.787	9.1	38.71 (28.80)	44.59 (26.61)	0.039	8.1
Legumes (g/day) (median [IQR])	36.65 [21.13, 40.97]	36.65 [25.33, 40.97]	0.838	9.1	38.57 [24.92, 43.13]	38.57 [26.67, 51.43]	0.002	8.1
Nuts (g/day) (mean (SD))	13.36 (19.82)	12.68 (19.13)	0.792	8.7	13.97 (19.31)	9.92 (12.99)	0.025	8.1
Nuts (g/day) (median [IQR])	6.43 [0.98, 17.14]	6.43 [2.46, 17.14]	0.635	8.7	6.43 [2.46, 19.10]	4.10 [0.98, 15.00]	0.024	8.1

Dairy and desserts (g/day) (mean (SD))	17.48 (30.09)	20.90 (39.09)	0.410	8.7	27.78 (48.77)	22.44 (26.87)	0.230	8.1
Dairy and desserts (g/day) (median [IQR])	7.02 [2.08, 20.61]	6.43 [2.83, 26.97]	0.351	8.7	11.91 [3.28, 31.36]	12.00 [3.71, 30.78]	0.781	8.1
Cheese (g/day) (mean (SD))	26.26 (30.37)	26.34 (29.93)	0.985	8.7	22.52 (33.06)	19.97 (21.57)	0.407	8.1
Cheese (g/day) (median [IQR])	17.36 [6.73, 34.07]	17.58 [7.28, 32.81]	0.882	8.7	15.00 [6.43, 27.91]	14.17 [6.43, 26.53]	0.332	8.1
Milk and yogurt (g/day) (mean (SD))	276.50 (174.65)	221.20 (163.19)	0.016	9.7	261.73 (178.02)	232.36 (167.43)	0.099	9.3
Milk and yogurt (g/day) (median [IQR])	228.55 [160.90, 369.00]	205.17 [88.80, 352.11]	0.018	9.7	226.34 [137.04, 388.05]	225.00 [101.83, 356.29]	0.114	9.3
Caloric beverages (g/day) (mean (SD))	77.00 (153.02)	79.31 (119.55)	0.906	9.1	127.77 (188.27)	137.96 (200.87)	0.600	9.0
Caloric beverages (g/day) (median [IQR])	19.68 [0.00, 66.62]	19.68 [0.00, 150.00]	0.851	9.1	53.50 [5.90, 164.20]	39.44 [0.00, 186.12]	0.827	9.0
Alcoholic beverages (g/day) (mean (SD))	83.51 (154.16)	85.76 (129.22)	0.910	8.7	212.68 (229.95)	269.63 (254.18)	0.018	9.5
Alcoholic beverages (g/day) (median [IQR])	20.34 [0.00, 99.08]	29.23 [0.00, 121.40]	0.622	8.7	142.39 [34.11, 306.71]	220.06 [65.33, 354.14]	0.021	9.5

Supplemental Table 2. Comparison across the original CRC Screening data and the data after imputation.

Original data: Variable [Missing]	**Original data: Total (749)**	**Imputed data: Total (749)**	**Original data: Cancer (224)**	**Imputed data: Cancer (224)**	**Original data: No cancer (525)**	**Imputed data: No cancer (525)**
Number of siblings (Mean/SD) [1]	2.94 (+-2.27)	2.94 (+-2.27)	2.7 (+-2.04)	2.7 (+-2.03)	3.04 (+-2.35)	3.04 (+-2.35)
Number of children (Mean/SD) [1]	1.82 (+-0.999)	1.82 (+-0.999)	1.83 (+-1.17)	1.83 (+-1.17)	1.82 (+-0.92)	1.82 (+-0.92)
Occupation [1]						
- Working	328 (43.8%)	329 (43.9%)	89 (39.9%)	90 (40.2%)	239 (45.5%)	239 (45.5%)
- Unemployed	63 (8.41%)	63 (8.41%)	22 (9.87%)	22 (9.82%)	41 (7.81%)	41 (7.81%)
- Housewife or domestic worker	72 (9.61%)	72 (9.61%)	21 (9.42%)	21 (9.38%)	51 (9.71%)	51 (9.71%)
- Retired	285 (38.1%)	285 (38.1%)	91 (40.8%)	91 (40.6%)	194 (37%)	194 (37%)
Weight (Mean/SD) [0]	77.8 (+-15.3)	77.8 (+-15.3)	79.3 (+-15.2)	79.3 (+-15.2)	77.1 (+-15.4)	77.1 (+-15.4)
High cholesterol [1]						
- No	462 (61.7%)	463 (61.8%)	127 (56.7%)	127 (56.7%)	335 (63.9%)	336 (64%)
- Yes	286 (38.2%)	286 (38.2%)	97 (43.3%)	97 (43.3%)	189 (36.1%)	189 (36%)
Osteoporosis [1]	40 (5.34%)	40 (5.34%)	11 (4.91%)	11 (4.91%)	29 (5.52%)	29 (5.52%)
Migraine [1]	67 (8.95%)	67 (8.95%)	19 (8.48%)	19 (8.48%)	48 (9.14%)	48 (9.14%)
Celiac disease [2]	2 (0.267%)	2 (0.267%)	0 (0%)	0 (0%)	2 (0.381%)	2 (0.381%)
Dyspepsia [2]	23 (3.07%)	24 (3.2%)	7 (3.12%)	7 (3.12%)	16 (3.05%)	17 (3.24%)
Fibromyalgia [2]	27 (3.6%)	27 (3.6%)	5 (2.23%)	5 (2.23%)	22 (4.19%)	22 (4.19%)
Schizophrenia [3]	3 (0.401%)	3 (0.401%)	0 (0%)	0 (0%)	3 (0.571%)	3 (0.571%)
Social class of parents [2]						
- Upper social class	5 (0.668%)	5 (0.668%)	2 (0.893%)	2 (0.893%)	3 (0.574%)	3 (0.571%)

- Middle social class	431 (57.5%)	433 (57.8%)	120 (53.6%)	120 (53.6%)	311 (59.5%)	313 (59.6%)
- Lower social class	311 (41.5%)	311 (41.5%)	102 (45.5%)	102 (45.5%)	209 (40%)	209 (39.8%)
Depression [4]	152 (20.3%)	153 (20.4%)	49 (21.9%)	49 (21.9%)	103 (19.6%)	104 (19.8%)
METs hours per week (Mean/SD) [6]	24.2 (+-24.7)	23.8 (+-24.6)	23.2 (+-24.7)	22.9 (+-24.6)	24.7 (+-24.7)	24.2 (+-24.6)
METs hours per week walking (Mean/SD) [1]	11.8 (+-17.3)	11.6 (+-17.2)	12.2 (+-17.3)	12 (+-17.2)	11.6 (+-17.4)	11.4 (+-17.2)
Maximum weight (Mean/SD) [4]	82.7 (+-16.6)	82.6 (+-16.6)	84.5 (+-16)	84.4 (+-16)	81.9 (+-16.8)	81.9 (+-16.8)
Age at maximum weight (Mean/SD) [6]	55.2 (+-14.1)	55.2 (+-14)	56.1 (+-14.7)	56.1 (+-14.7)	54.9 (+-13.8)	54.9 (+-13.7)
Weight 1 year ago (Mean/SD) [11]	77.9 (+-15.8)	77.8 (+-15.9)	79.6 (+-15.4)	79.5 (+-15.5)	77.2 (+-16)	77.1 (+-16)
Physical activity at work [10]						
- Sedentary	97 (13%)	98 (13.1%)	28 (12.9%)	28 (12.5%)	69 (13.2%)	70 (13.3%)
- Slightly active	112 (15%)	115 (15.4%)	24 (11.1%)	27 (12.1%)	88 (16.9%)	88 (16.8%)
- Moderately active	199 (26.6%)	201 (26.8%)	64 (29.5%)	65 (29%)	135 (25.9%)	136 (25.9%)
- Fairly active	237 (31.6%)	240 (32%)	64 (29.5%)	67 (29.9%)	173 (33.1%)	173 (33%)
- Very active	94 (12.6%)	95 (12.7%)	37 (17.1%)	37 (16.5%)	57 (10.9%)	58 (11%)
Waist circumference (Mean/SD) [75]	96.2 (+-12.8)	96 (+-12.6)	98.5 (+-12.3)	98.1 (+-12.2)	95.2 (+-13)	95.1 (+-12.6)
Hip circumference (Mean/SD) [103]	103 (+-10.4)	103 (+-9.97)	104 (+-11)	104 (+-10.5)	103 (+-10.1)	103 (+-9.72)
During the past month, how often have you thought about your chances of getting cancer? [80]						

- Rarely or never	254 (33.9%)	291 (38.9%)	45 (31.2%)	82 (36.6%)	209 (39.8%)	209 (39.8%)
- Sometimes, often, or almost all the time	415 (55.4%)	458 (61.1%)	99 (68.8%)	142 (63.4%)	316 (60.2%)	316 (60.2%)
During the past month, has thinking about the possibility of developing cancer affected your mood? [81]						
- Rarely or never	387 (51.7%)	424 (56.6%)	77 (53.8%)	114 (50.9%)	310 (59%)	310 (59%)
- Sometimes, often, or almost all the time	281 (37.5%)	325 (43.4%)	66 (46.2%)	110 (49.1%)	215 (41%)	215 (41%)
To what extent do you worry about the possibility of developing cancer one day? [188]						
- Not at all	184 (24.6%)	208 (27.8%)	44 (38.9%)	64 (28.6%)	140 (31.2%)	144 (27.4%)
- A little	281 (37.5%)	392 (52.3%)	53 (46.9%)	123 (54.9%)	228 (50.9%)	269 (51.2%)
- Quite a bit or a great deal	96 (12.8%)	149 (19.9%)	16 (14.2%)	37 (16.5%)	80 (17.9%)	112 (21.3%)
How often do you worry about the possibility of developing cancer? [81]						
- Frequently or Constantly	55 (7.34%)	69 (9.21%)	14 (9.72%)	28 (12.5%)	41 (7.82%)	41 (7.81%)
- Never or rarely	264 (35.2%)	289 (38.6%)	61 (42.4%)	86 (38.4%)	203 (38.7%)	203 (38.7%)
- Occasionally	349 (46.6%)	391 (52.2%)	69 (47.9%)	110 (49.1%)	280 (53.4%)	281 (53.5%)

Is being worried about developing cancer an important issue for you? [81]						
- No	249 (33.2%)	287 (38.3%)	59 (41%)	97 (43.3%)	190 (36.3%)	190 (36.2%)
- Yes	419 (55.9%)	462 (61.7%)	85 (59%)	127 (56.7%)	334 (63.7%)	335 (63.8%)
During the past month, has thinking about the possibility of developing cancer affected your ability to carry out your daily activities? [82]						
- Rarely or never	439 (58.6%)	497 (66.4%)	93 (65%)	150 (67%)	346 (66%)	347 (66.1%)
- Sometimes, often, or almost all the time	228 (30.4%)	252 (33.6%)	50 (35%)	74 (33%)	178 (34%)	178 (33.9%)
Willing to change the lifestyle to reduce colon cancer risk [104]	624 (83.3%)	692 (92.4%)	131 (58.5%)	187 (83.5%)	493 (93.9%)	505 (96.2%)
If you were obese, would you lose weight? [94]	503 (67.2%)	536 (71.6%)	103 (46%)	135 (60.3%)	400 (76.2%)	401 (76.4%)
If you did little exercise: would you do more exercise on a regular basis? [94]	603 (80.5%)	642 (85.7%)	126 (56.2%)	160 (71.4%)	477 (90.9%)	482 (91.8%)
If you were to eat a meat-heavy diet: would you eat less meat? [94]	566 (75.6%)	589 (78.6%)	116 (51.8%)	137 (61.2%)	450 (85.7%)	452 (86.1%)
If you were to eat a diet low in vegetables: would you eat more vegetables? [94]	576 (76.9%)	604 (80.6%)	123 (54.9%)	150 (67%)	453 (86.3%)	454 (86.5%)

If you were a heavy drinker, would you reduce your alcohol consumption? [95]	265 (35.4%)	286 (38.2%)	73 (32.6%)	90 (40.2%)	192 (36.6%)	196 (37.3%)
If you were a smoker, would you quit smoking? [101]	236 (31.5%)	265 (35.4%)	68 (30.4%)	90 (40.2%)	168 (32%)	175 (33.3%)
Are early detection programs useful? [85]	608 (81.2%)	667 (89.1%)	130 (58%)	187 (83.5%)	478 (91%)	480 (91.4%)
Willing to participate again? [114]	632 (84.4%)	727 (97.1%)	130 (58%)	205 (91.5%)	502 (95.6%)	522 (99.4%)
Current smoker [1]	178 (23.8%)	178 (23.8%)	82 (36.6%)	82 (36.6%)	96 (18.3%)	96 (18.3%)
Total number of cigarettes smoked in a lifetime (Mean/SD) [30]	140,685 (+-178,965)	141,370 (+-177,595)	190,871 (+-210,234)	189,440 (+-206,629)	119,980 (+-160,044)	120,860 (+-159,509)
Pack years (Mean/SD) [30]	19.3 (+-24.5)	19 (+-24.2)	26.1 (+-28.8)	25.5 (+-28.2)	16.4 (+-21.9)	16.3 (+-21.8)
Years of smoking (Mean/SD) [15]	19.8 (+-17.7)	19.9 (+-17.6)	26 (+-18.2)	25.7 (+-18.1)	17.2 (+-16.9)	17.4 (+-16.9)
Reason for last menstruation [38]						
- Natural menopause	237 (31.6%)	252 (33.6%)	50 (23.6%)	56 (25%)	187 (37.5%)	196 (37.3%)
- Removal of the uterus and ovaries	14 (1.87%)	22 (2.94%)	6 (2.83%)	9 (4.02%)	8 (1.6%)	13 (2.48%)
- Removal of the uterus only	25 (3.34%)	31 (4.14%)	6 (2.83%)	7 (3.12%)	19 (3.81%)	24 (4.57%)
- Removal of the ovaries only	1 (0.134%)	2 (0.267%)	0 (0%)	0 (0%)	1 (0.2%)	2 (0.381%)
- Still has periods	20 (2.67%)	27 (3.6%)	1 (0.472%)	2 (0.893%)	19 (3.81%)	25 (4.76%)
- Other causes	5 (0.668%)	6 (0.801%)	1 (0.472%)	2 (0.893%)	4 (0.802%)	4 (0.762%)
- Not applicable	409 (54.6%)	409 (54.6%)	148 (69.8%)	148 (66.1%)	261 (52.3%)	261 (49.7%)

Number of cigarettes on average (Mean/SD) [1]	3.35 (+-7.74)	3.3 (+-7.7)	4.96 (+-8.81)	4.87 (+-8.75)	2.66 (+-7.14)	2.64 (+-7.11)
Total energy (kcal/day) (Mean/SD) [23]	1,819 (+-620)	1,815 (+-612)	1,850 (+-616)	1,835 (+-588)	1,807 (+-622)	1,807 (+-622)
Total protein (g/day) (Mean/SD) [23]	80.4 (+-26.5)	80.2 (+-26.2)	81.6 (+-26.4)	80.8 (+-25.2)	79.9 (+-26.6)	79.9 (+-26.6)
Total carbohydrates (g/day) (Mean/SD) [23]	177 (+-68.2)	177 (+-67.3)	177 (+-65)	175 (+-62)	178 (+-69.5)	178 (+-69.5)
Total fats (g/day) (Mean/SD) [23]	77.6 (+-31.1)	77.5 (+-30.7)	77.7 (+-30.9)	77.4 (+-29.5)	77.5 (+-31.3)	77.5 (+-31.3)
Total fiber (g/day) (Mean/SD) [23]	20 (+-9.2)	19.9 (+-9.09)	19.1 (+-7.87)	18.9 (+-7.55)	20.3 (+-9.64)	20.3 (+-9.64)
Total ethanol (g/day) (Mean/SD) [23]	12.1 (+-17.2)	12 (+-17)	16 (+-20.1)	15.1 (+-19.3)	10.6 (+-15.7)	10.6 (+-15.7)
White meat (g/day) (Mean/SD) [23]	28.4 (+-21.8)	28.2 (+-21.5)	28.2 (+-20.9)	27.5 (+-20)	28.5 (+-22.1)	28.5 (+-22.1)
Cured and processed meat (g/day) (Mean/SD) [23]	38.9 (+-27)	38.7 (+-26.6)	42.4 (+-28.6)	41.6 (+-27.4)	37.5 (+-26.2)	37.5 (+-26.2)
All meat (g/day) (Mean/SD) [23]	95.8 (+-53)	95.6 (+-52.3)	102 (+-57)	101 (+-54.5)	93.3 (+-51.3)	93.3 (+-51.3)
Red meat (g/day) (Mean/SD) [23]	27.5 (+-25.5)	27.4 (+-25.2)	30.2 (+-27.5)	29.6 (+-26.3)	26.5 (+-24.7)	26.5 (+-24.7)
White fish (g/day) (Mean/SD) [23]	15.8 (+-13.2)	15.7 (+-13.1)	16.6 (+-13.9)	16.1 (+-13.3)	15.5 (+-13)	15.5 (+-13)
Blue fish (g/day) (Mean/SD) [23]	14.4 (+-12.7)	14.2 (+-12.6)	14.8 (+-12.1)	14.1 (+-11.7)	14.2 (+-13)	14.2 (+-13)
Fruits (g/day) (Mean/SD) [29]	241 (+-173)	242 (+-173)	219 (+-153)	216 (+-147)	250 (+-179)	254 (+-182)
Vegetables (g/day) (Mean/SD) [24]	175 (+-110)	174 (+-110)	172 (+-116)	169 (+-112)	176 (+-108)	177 (+-109)
Legumes (g/day) (Mean/SD) [24]	39.8 (+-32)	39.4 (+-31.6)	42.8 (+-28.6)	41.3 (+-27.7)	38.6 (+-33.1)	38.6 (+-33.1)

Nuts (g/day) (Mean/SD) [23]	13.3 (+-19.2)	13 (+-18.9)	10.9 (+-15.4)	10.3 (+-14.7)	14.2 (+-20.4)	14.2 (+-20.4)
Dairy and desserts (g/day) (Mean/SD) [23]	21.2 (+-36)	20.8 (+-35.5)	21.9 (+-31.4)	20.6 (+-30.1)	20.9 (+-37.6)	20.9 (+-37.6)
Cheese (g/day) (Mean/SD) [23]	25 (+-32.9)	24.7 (+-32.4)	22.1 (+-24.8)	21.4 (+-23.8)	26.1 (+-35.4)	26.1 (+-35.4)
Milk and yogurt (g/day) (Mean/SD) [34]	259 (+-176)	254 (+-175)	229 (+-166)	217 (+-162)	270 (+-179)	270 (+-178)
Caloric beverages (g/day) (Mean/SD) [30]	105 (+-177)	104 (+-175)	118 (+-179)	110 (+-172)	99.8 (+-176)	101 (+-176)
Alcoholic beverages (g/day) (Mean/SD) [28]	164 (+-219)	163 (+-218)	207 (+-236)	195 (+-228)	148 (+-210)	150 (+-212)
Circulatory problems [2]	70 (9.35%)	70 (9.35%)	20 (8.93%)	20 (8.93%)	50 (9.52%)	50 (9.52%)
Arthritis [2]	159 (21.2%)	159 (21.2%)	57 (25.4%)	57 (25.4%)	102 (19.4%)	102 (19.4%)
Anemia [2]	43 (5.74%)	43 (5.74%)	11 (4.91%)	11 (4.91%)	32 (6.1%)	32 (6.1%)
Diverticulitis [3]	9 (1.2%)	9 (1.2%)	4 (1.79%)	4 (1.79%)	5 (0.952%)	5 (0.952%)
Prostate disease [2]	83 (11.1%)	83 (11.1%)	26 (11.6%)	26 (11.6%)	57 (10.9%)	57 (10.9%)
Heartburn [4]	256 (34.2%)	256 (34.2%)	69 (30.8%)	69 (30.8%)	187 (35.6%)	187 (35.6%)
Laxative use [9]	133 (17.8%)	136 (18.2%)	29 (12.9%)	30 (13.4%)	104 (19.8%)	106 (20.2%)
Anti-inflammatory medication [43]	170 (22.7%)	178 (23.8%)	51 (22.8%)	54 (24.1%)	119 (22.7%)	124 (23.6%)
Menopause treatment [11]	40 (5.34%)	40 (5.34%)	11 (4.91%)	11 (4.91%)	29 (5.52%)	29 (5.52%)
Use contraceptive [2]	230 (30.7%)	231 (30.8%)	47 (21%)	48 (21.4%)	183 (34.9%)	183 (34.9%)
Age at last menstruation [42]						

- 33 - 46	73 (9.75%)	82 (10.9%)	21 (9.91%)	24 (10.7%)	52 (10.5%)	58 (11%)
- 46 - 50	116 (15.5%)	138 (18.4%)	23 (10.8%)	29 (12.9%)	93 (18.8%)	109 (20.8%)
- 50 - 52	37 (4.94%)	43 (5.74%)	8 (3.77%)	9 (4.02%)	29 (5.86%)	34 (6.48%)
- 52 - 60	52 (6.94%)	54 (7.21%)	11 (5.19%)	12 (5.36%)	41 (8.28%)	42 (8%)
- Non applicable	409 (54.6%)	409 (54.6%)	148 (69.8%)	148 (66.1%)	261 (52.7%)	261 (49.7%)
- Still has periods	20 (2.67%)	23 (3.07%)	1 (0.472%)	2 (0.893%)	19 (3.84%)	21 (4%)
Average lifetime intensity in cigarettes/year (Mean/SD) [24]	4,597 (+-5,362)	4,557 (+-5,309)	5,510 (+-5,772)	5,401 (+-5,699)	4,218 (+-5,140)	4,197 (+-5,097)
In your lifetime, have you ever smoked? 'YES' means at least 100 cigarettes or 360 grams of tobacco in your lifetime. [0]	491 (65.6%)	491 (65.6%)	168 (75%)	168 (75%)	323 (61.5%)	323 (61.5%)
Age at smoking initiation [1]						
- 8 - 15	152 (20.3%)	152 (20.3%)	54 (24.2%)	54 (24.1%)	98 (18.7%)	98 (18.7%)
- 15 - 17	116 (15.5%)	117 (15.6%)	41 (18.4%)	42 (18.8%)	75 (14.3%)	75 (14.3%)
- 17 - 19	106 (14.2%)	106 (14.2%)	30 (13.5%)	30 (13.4%)	76 (14.5%)	76 (14.5%)
- 19 - 54	115 (15.4%)	115 (15.4%)	42 (18.8%)	42 (18.8%)	73 (13.9%)	73 (13.9%)
- Never smoked	259 (34.6%)	259 (34.6%)	56 (25.1%)	56 (25%)	203 (38.7%)	203 (38.7%)
Smoking status [1]						
- Never	258 (34.4%)	259 (34.6%)	56 (25%)	56 (25%)	202 (38.5%)	203 (38.7%)
- Ex-Smoker	312 (41.7%)	312 (41.7%)	86 (38.4%)	86 (38.4%)	226 (43.1%)	226 (43%)

- Smoker	178 (23.8%)	178 (23.8%)	82 (36.6%)	82 (36.6%)	96 (18.3%)	96 (18.3%)
Current frequency of smoking [1]						
- Day	174 (23.2%)	174 (23.2%)	80 (35.7%)	80 (35.7%)	94 (17.9%)	94 (17.9%)
- Week	4 (0.534%)	4 (0.534%)	2 (0.893%)	2 (0.893%)	2 (0.382%)	2 (0.381%)
- Former smoker	312 (41.7%)	312 (41.7%)	86 (38.4%)	86 (38.4%)	226 (43.1%)	226 (43%)
- Never	258 (34.4%)	259 (34.6%)	56 (25%)	56 (25%)	202 (38.5%)	203 (38.7%)
Passive smoker [175]	171 (22.8%)	197 (26.3%)	58 (25.9%)	66 (29.5%)	113 (21.5%)	131 (25%)
Average annual cigarettes during the time smoked [186]						
- 0 - 3652	106 (14.2%)	142 (19%)	22 (15.8%)	33 (14.7%)	84 (19.8%)	109 (20.8%)
- 3652 - 7305	106 (14.2%)	125 (16.7%)	37 (26.6%)	43 (19.2%)	69 (16.3%)	82 (15.6%)
- 7305 - 29220	92 (12.3%)	223 (29.8%)	24 (17.3%)	92 (41.1%)	68 (16%)	131 (25%)
- Never smoked	259 (34.6%)	259 (34.6%)	56 (40.3%)	56 (25%)	203 (47.9%)	203 (38.7%)
Annual Average of cigarettes per day during the time smoked [186]						
- 0 - 10	106 (14.2%)	144 (19.2%)	22 (15.8%)	33 (14.7%)	84 (19.8%)	111 (21.1%)
- 10 - 20	106 (14.2%)	129 (17.2%)	37 (26.6%)	47 (21%)	69 (16.3%)	82 (15.6%)
- 20 - 30	38 (5.07%)	66 (8.81%)	7 (5.04%)	18 (8.04%)	31 (7.31%)	48 (9.14%)
- 30 - 80	54 (7.21%)	151 (20.2%)	17 (12.2%)	70 (31.2%)	37 (8.73%)	81 (15.4%)

- Never smoked	259 (34.6%)	259 (34.6%)	56 (40.3%)	56 (25%)	203 (47.9%)	203 (38.7%)
Waist circumference (Mean/SD) [75]	96.2 (+-12.8)	96 (+-12.6)	98.5 (+-12.3)	98.1 (+-12.2)	95.2 (+-13)	95.1 (+-12.6)
Hip circumference (Mean/SD) [103]	103 (+-10.4)	103 (+-9.97)	104 (+-11)	104 (+-10.5)	103 (+-10.1)	103 (+-9.72)
Age at first menstruation [5]						
- 11 - 13	134 (17.9%)	139 (18.6%)	33 (14.8%)	34 (15.2%)	101 (19.4%)	105 (20%)
- 13 - 14	79 (10.5%)	79 (10.5%)	10 (4.48%)	10 (4.46%)	69 (13.2%)	69 (13.1%)
- 14 - 18	28 (3.74%)	28 (3.74%)	5 (2.24%)	5 (2.23%)	23 (4.41%)	23 (4.38%)
- 8 - 11	94 (12.6%)	94 (12.6%)	27 (12.1%)	27 (12.1%)	67 (12.9%)	67 (12.8%)
- Non applicable	409 (54.6%)	409 (54.6%)	148 (66.4%)	148 (66.1%)	261 (50.1%)	261 (49.7%)
Menstruation status [1]						
- Still has periods	20 (2.67%)	20 (2.67%)	1 (0.448%)	1 (0.446%)	19 (3.62%)	19 (3.62%)
- Is menopausal	37 (4.94%)	37 (4.94%)	11 (4.93%)	11 (4.91%)	26 (4.95%)	26 (4.95%)
- No longer has periods/postmenopausa	282 (37.7%)	283 (37.8%)	63 (28.3%)	64 (28.6%)	219 (41.7%)	219 (41.7%)
- Not applicable	409 (54.6%)	409 (54.6%)	148 (66.4%)	148 (66.1%)	261 (49.7%)	261 (49.7%)

Supplemental Table 3. Performance metrics for the best performing models for the subgroup of females for CRC data.

Learner	Accuracy	AUC	PRAUC	F1	Precision	Recall	Macro F1
naive_bayes	0.753 (+-0.075)	0.671 (+-0.116)	0.413 (+-0.149)	0.493 (+-0.132)	0.667 (+-0.091)	0.679 (+-0.102)	0.664 (+-0.089)
lda	0.727 (+-0.059)	0.676 (+-0.086)	0.45 (+-0.141)	0.456 (+-0.115)	0.631 (+-0.071)	0.651 (+-0.082)	0.636 (+-0.075)
glmnet	0.7 (+-0.053)	0.657 (+-0.096)	0.429 (+-0.176)	0.416 (+-0.099)	0.605 (+-0.058)	0.625 (+-0.074)	0.607 (+-0.062)
ranger	0.75 (+-0.075)	0.674 (+-0.101)	0.413 (+-0.128)	0.355 (+-0.186)	0.621 (+-0.115)	0.598 (+-0.11)	0.599 (+-0.114)

Supplemental Table 4. Performance metrics for the best performing models for the subgroup of males for CRC data.

Learner	Accuracy	AUC	PRAUC	F1	Precision	Recall	Macro F1
naive_bayes	0.719 (+-0.079)	0.72 (+-0.095)	0.651 (+-0.134)	0.577 (+-0.128)	0.697 (+-0.091)	0.681 (+-0.091)	0.683 (+-0.091)
lda	0.665 (+-0.06)	0.659 (+-0.099)	0.584 (+-0.118)	0.505 (+-0.097)	0.638 (+-0.063)	0.626 (+-0.066)	0.624 (+-0.066)
glmnet	0.653 (+-0.039)	0.636 (+-0.098)	0.557 (+-0.123)	0.496 (+-0.092)	0.622 (+-0.054)	0.617 (+-0.056)	0.614 (+-0.053)
nnet	0.633 (+-0.084)	0.647 (+-0.101)	0.559 (+-0.118)	0.483 (+-0.128)	0.605 (+-0.091)	0.602 (+-0.095)	0.598 (+-0.092)

Supplemental Table 5. Comparison across the original COSMOS data, the data after missing imputation, and the data after applying SMOTE (Synthetic Minority Over-sampling Technique) for class imbalance.

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
Female [0]	984 (36.6%)	984 (36.6%)	171 (38.9%)	42 (38.2%)	42 (38.2%)	79 (47.9%)	942 (36.5%)	942 (36.5%)	92 (33.5%)
Age (Mean/SD) [0]	62.1 (+-4.94)	61.6 (+-4.94)	62 (+-5.09)	63.6 (+-5.77)	63.2 (+-5.76)	62.3 (+-5.48)	62.1 (+-4.9)	61.6 (+-4.89)	61.8 (+-4.83)
BMI (Mean/SD) [7]	25.7 (+-5.61)	25.2 (+-5.61)	25 (+-4.55)	25 (+-5.38)	24.5 (+-5.4)	24.8 (+-5.2)	25.7 (+-5.62)	25.3 (+-5.62)	25.2 (+-4.11)
Frequency of usual consum									

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
ption of a portion of raw or cooked vegetables, salad included (150 g) [72]									
- Rarely (never/ 1-2 times a month)	101 (3.75%)	101 (3.75%)	8 (1.82%)	5 (4.85%)	5 (4.55%)	5 (3.03%)	96 (3.82%)	96 (3.72%)	3 (1.09%)
- Once a week	235 (8.74%)	236 (8.77%)	51 (11.6%)	9 (8.74%)	9 (8.18%)	23 (13.9%)	226 (8.99%)	227 (8.8%)	28 (10.2%)
- 2-3 times a week	829 (30.8%)	870 (32.3%)	144 (32.7%)	34 (33%)	36 (32.7%)	54 (32.7%)	795 (31.6%)	834 (32.3%)	90 (32.7%)
- Every day	1,104 (41%)	1,131 (42%)	180 (40.9%)	41 (39.8%)	45 (40.9%)	61 (37%)	1,063 (42.3%)	1,086 (42.1%)	119 (43.3%)
- Several times a day	349 (13%)	352 (13.1%)	57 (13%)	14 (13.6%)	15 (13.6%)	22 (13.3%)	335 (13.3%)	337 (13.1%)	35 (12.7%)
Frequency of usual consumption of a portion of fresh fruit (all types - 150 g) [93]									
- Rarely (never/ 1-2	164 (6.1%)	166 (6.17%)	28 (6.36%)	9 (8.82%)	9 (8.18%)	9 (5.45%)	155 (6.21%)	157 (6.09%)	19 (6.91%)

Original data: Variable [Missing]	**Original data: Total (2,690)**	**Imputed data: Total (2,690)**	**Balanced data: Total (440)**	**Original data: Lung cancer (110)**	**Imputed data: Lung cancer (110)**	**Balanced data: Lung cancer (165)**	**Original data: No cancer (2,580)**	**Imputed data: No cancer (2,580)**	**Balanced data: No cancer (275)**
times a month)									
- Once a week	190 (7.06%)	193 (7.17%)	25 (5.68%)	4 (3.92%)	4 (3.64%)	7 (4.24%)	186 (7.45%)	189 (7.33%)	18 (6.55%)
- 2-3 times a week	569 (21.2%)	585 (21.7%)	99 (22.5%)	22 (21.6%)	22 (20%)	35 (21.2%)	547 (21.9%)	563 (21.8%)	64 (23.3%)
- Every day	1,244 (46.2%)	1,311 (48.7%)	218 (49.5%)	45 (44.1%)	53 (48.2%)	88 (53.3%)	1,199 (48.1%)	1,258 (48.8%)	130 (47.3%)
- Several times a day	430 (16%)	435 (16.2%)	70 (15.9%)	22 (21.6%)	22 (20%)	26 (15.8%)	408 (16.4%)	413 (16%)	44 (16%)
Frequency of usual consumption of a portion of white meat (chicken, turkey, rabbit - 100 g) [137]									
- Rarely (never/ 1-2 times a month)	394 (14.6%)	400 (14.9%)	53 (12%)	15 (15.3%)	16 (14.5%)	20 (12.1%)	379 (15.4%)	384 (14.9%)	33 (12%)
- Once a week	903 (33.6%)	942 (35%)	145 (33%)	39 (39.8%)	42 (38.2%)	48 (29.1%)	864 (35.2%)	900 (34.9%)	97 (35.3%)
- 2-3 times a week	1,187 (44.1%)	1,278 (47.5%)	234 (53.2%)	42 (42.9%)	50 (45.5%)	95 (57.6%)	1,145 (46.6%)	1,228 (47.6%)	139 (50.5%)

Original data: Variable [Missing]	**Original data: Total (2,690)**	**Imputed data: Total (2,690)**	**Balanced data: Total (440)**	**Original data: Lung cancer (110)**	**Imputed data: Lung cancer (110)**	**Balanced data: Lung cancer (165)**	**Original data: No cancer (2,580)**	**Imputed data: No cancer (2,580)**	**Balanced data: No cancer (275)**
- Every day	60 (2.23%)	60 (2.23%)	8 (1.82%)	2 (2.04%)	2 (1.82%)	2 (1.21%)	58 (2.36%)	58 (2.25%)	6 (2.18%)
- Several times a day	9 (0.335%)	10 (0.372%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	9 (0.367%)	10 (0.388%)	0 (0%)
Frequency of usual consumption of a portion of red meat (beef, veal, pork - 100 g) [167]									
- Rarely (never/ 1-2 times a month)	507 (18.8%)	523 (19.4%)	62 (14.1%)	18 (17.8%)	19 (17.3%)	23 (13.9%)	489 (20.2%)	504 (19.5%)	39 (14.2%)
- Once a week	1,055 (39.2%)	1,118 (41.6%)	192 (43.6%)	44 (43.6%)	47 (42.7%)	72 (43.6%)	1,011 (41.7%)	1,071 (41.5%)	120 (43.6%)
- 2-3 times a week	920 (34.2%)	1,005 (37.4%)	179 (40.7%)	38 (37.6%)	43 (39.1%)	68 (41.2%)	882 (36.4%)	962 (37.3%)	111 (40.4%)
- Every day	40 (1.49%)	41 (1.52%)	7 (1.59%)	1 (0.99%)	1 (0.909%)	2 (1.21%)	39 (1.61%)	40 (1.55%)	5 (1.82%)
- Several times a day	1 (0.037%)	3 (0.112%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (0.041%)	3 (0.116%)	0 (0%)
Frequency of usual									

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
consumption of a portion of cold cuts, cured meats, and sausages (e.g., ham, salami, bresaola/dried beef, sausages, etc. - 50 g) [107]									
- Rarely (never/ 1-2 times a month)	496 (18.4%)	505 (18.8%)	61 (13.9%)	16 (15.5%)	16 (14.5%)	18 (10.9%)	480 (19.4%)	489 (19%)	43 (15.6%)
- Once a week	893 (33.2%)	932 (34.6%)	140 (31.8%)	28 (27.2%)	31 (28.2%)	37 (22.4%)	865 (34.9%)	901 (34.9%)	103 (37.5%)
- 2-3 times a week	1,065 (39.6%)	1,122 (41.7%)	213 (48.4%)	53 (51.5%)	57 (51.8%)	90 (54.5%)	1,012 (40.8%)	1,065 (41.3%)	123 (44.7%)
- Every day	119 (4.42%)	120 (4.46%)	26 (5.91%)	6 (5.83%)	6 (5.45%)	20 (12.1%)	113 (4.56%)	114 (4.42%)	6 (2.18%)
- Several times a day	10 (0.372%)	11 (0.409%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	10 (0.403%)	11 (0.426%)	0 (0%)
Alcohol consumption (e.g., glass of wine, beer,									

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
liquor) [57]									
- Never	856 (31.8%)	870 (32.3%)	134 (30.5%)	44 (41.1%)	44 (40%)	50 (30.3%)	812 (32.1%)	826 (32%)	84 (30.5%)
- ≤4 glasses/ week	39 (1.45%)	40 (1.49%)	4 (0.909%)	0 (0%)	0 (0%)	0 (0%)	39 (1.54%)	40 (1.55%)	4 (1.45%)
- 1-2 glasses/ day	1,205 (44.8%)	1,244 (46.2%)	218 (49.5%)	40 (37.4%)	43 (39.1%)	86 (52.1%)	1,165 (46.1%)	1,201 (46.6%)	132 (48%)
- 3-5 glasses/ day	474 (17.6%)	477 (17.7%)	76 (17.3%)	21 (19.6%)	21 (19.1%)	27 (16.4%)	453 (17.9%)	456 (17.7%)	49 (17.8%)
- >5 glasses/ day	59 (2.19%)	59 (2.19%)	8 (1.82%)	2 (1.87%)	2 (1.82%)	2 (1.21%)	57 (2.26%)	57 (2.21%)	6 (2.18%)
Have you had any chest diagnostic tests performed in the last year? [46]	514 (19.1%)	520 (19.3%)	81 (18.4%)	20 (18.2%)	21 (19.1%)	32 (19.4%)	494 (19.1%)	499 (19.3%)	49 (17.8%)
Chronic bronchitis [87]	452 (16.8%)	460 (17.1%)	107 (24.3%)	29 (26.4%)	30 (27.3%)	57 (34.5%)	423 (16.4%)	430 (16.7%)	50 (18.2%)
Pneumonia [27]	409 (15.2%)	413 (15.4%)	77 (17.5%)	22 (20%)	22 (20%)	32 (19.4%)	387 (15%)	391 (15.2%)	45 (16.4%)
Tuberculosis [60]	48 (1.78%)	49 (1.82%)	6 (1.36%)	1 (0.909%)	1 (0.909%)	1 (0.606%)	47 (1.82%)	48 (1.86%)	5 (1.82%)
Pleurisy [48]	121 (4.5%)	122 (4.54%)	26 (5.91%)	7 (6.36%)	7 (6.36%)	12 (7.27%)	114 (4.42%)	115 (4.46%)	14 (5.09%)

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
Pneumothorax [78]	32 (1.19%)	32 (1.19%)	7 (1.59%)	1 (0.909%)	1 (0.909%)	1 (0.606%)	31 (1.2%)	31 (1.2%)	6 (2.18%)
Asthma [52]	140 (5.2%)	141 (5.24%)	17 (3.86%)	4 (3.64%)	4 (3.64%)	6 (3.64%)	136 (5.27%)	137 (5.31%)	11 (4%)
Other allergies [109]	447 (16.6%)	469 (17.4%)	60 (13.6%)	15 (13.6%)	16 (14.5%)	22 (13.3%)	432 (16.7%)	453 (17.6%)	38 (13.8%)
Cardiovascular diseases [107]	446 (16.6%)	460 (17.1%)	66 (15%)	22 (20%)	22 (20%)	31 (18.8%)	424 (16.4%)	438 (17%)	35 (12.7%)
Thyroid diseases [158]	312 (11.6%)	322 (12%)	61 (13.9%)	16 (14.5%)	16 (14.5%)	37 (22.4%)	296 (11.5%)	306 (11.9%)	24 (8.73%)
Other comorbidities [0]	422 (15.7%)	422 (15.7%)	80 (18.2%)	22 (20%)	22 (20%)	45 (27.3%)	400 (15.5%)	400 (15.5%)	35 (12.7%)
Are you currently undergoing drug therapy? [33]	1,825 (67.8%)	1,847 (68.7%)	318 (72.3%)	81 (73.6%)	82 (74.5%)	135 (81.8%)	1,744 (67.6%)	1,765 (68.4%)	183 (66.5%)
Family history of lung cancer [254]	721 (26.8%)	770 (28.6%)	142 (32.3%)	33 (30%)	36 (32.7%)	69 (41.8%)	688 (26.7%)	734 (28.4%)	73 (26.5%)
Family member with a history of lung cancer [256]									
- No family history	1,715 (63.8%)	1,930 (71.7%)	314 (71.4%)	65 (66.3%)	75 (68.2%)	112 (67.9%)	1,650 (70.6%)	1,855 (71.9%)	202 (73.5%)

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
- Father	384 (14.3%)	418 (15.5%)	76 (17.3%)	16 (16.3%)	18 (16.4%)	34 (20.6%)	368 (15.8%)	400 (15.5%)	42 (15.3%)
- Mother	84 (3.12%)	88 (3.27%)	10 (2.27%)	3 (3.06%)	3 (2.73%)	3 (1.82%)	81 (3.47%)	85 (3.29%)	7 (2.55%)
- Brother	80 (2.97%)	83 (3.09%)	10 (2.27%)	4 (4.08%)	4 (3.64%)	4 (2.42%)	76 (3.25%)	79 (3.06%)	6 (2.18%)
- Sister	28 (1.04%)	28 (1.04%)	7 (1.59%)	3 (3.06%)	3 (2.73%)	3 (1.82%)	25 (1.07%)	25 (0.969%)	4 (1.45%)
- Other	143 (5.32%)	143 (5.32%)	23 (5.23%)	7 (7.14%)	7 (6.36%)	9 (5.45%)	136 (5.82%)	136 (5.27%)	14 (5.09%)
Do you currently smoke? [9]									
- Yes	2,081 (77.4%)	2,085 (77.5%)	354 (80.5%)	87 (79.8%)	88 (80%)	139 (84.2%)	1,994 (77.5%)	1,997 (77.4%)	215 (78.2%)
- No, former smoker	600 (22.3%)	605 (22.5%)	86 (19.5%)	22 (20.2%)	22 (20%)	26 (15.8%)	578 (22.5%)	583 (22.6%)	60 (21.8%)
At what age did you start smoking? (Mean/SD) [12]	17.4 (+-3.9)	16.9 (+-4.03)	16.4 (+-3.39)	17.1 (+-3.7)	16.5 (+-3.79)	16.1 (+-3.63)	17.4 (+-3.91)	16.9 (+-4.04)	16.6 (+-3.23)
For how many years did you smoke in total?	41.4 (+-7.03)	41.1 (+-7.01)	41.9 (+-6.94)	44.1 (+-7.4)	43.8 (+-7.44)	43.4 (+-6.7)	41.3 (+-6.99)	41 (+-6.97)	41 (+-6.93)

Original data: Variable [Missing]	**Original data: Total (2,690)**	**Imputed data: Total (2,690)**	**Balanced data: Total (440)**	**Original data: Lung cancer (110)**	**Imputed data: Lung cancer (110)**	**Balanced data: Lung cancer (165)**	**Original data: No cancer (2,580)**	**Imputed data: No cancer (2,580)**	**Balanced data: No cancer (275)**
(Mean/ SD) [16]									
Pack/years (Mean/ SD) [39]	56.7 (+-219)	56.2 (+-217)	56.6 (+-175)	86.3 (+-350)	85.8 (+-349)	77.6 (+-285)	55.5 (+-211)	54.9 (+-210)	44 (+-18.1)
Type of cigarettes smoked [71]									
- Filtered	2,571 (95.6%)	2,642 (98.2%)	430 (97.7%)	106 (98.1%)	108 (98.2%)	159 (96.4%)	2,465 (98.2%)	2,534 (98.2%)	271 (98.5%)
- Unfiltered	48 (1.78%)	48 (1.78%)	10 (2.27%)	2 (1.85%)	2 (1.82%)	6 (3.64%)	46 (1.83%)	46 (1.78%)	4 (1.45%)
Have you ever smoked cigars? [209]									
- Yes	330 (12.3%)	358 (13.3%)	80 (18.2%)	19 (18.1%)	21 (19.1%)	45 (27.3%)	311 (13.1%)	337 (13.1%)	35 (12.7%)
- No	2,151 (80%)	2,332 (86.7%)	360 (81.8%)	86 (81.9%)	89 (80.9%)	120 (72.7%)	2,065 (86.9%)	2,243 (86.9%)	240 (87.3%)
Have you ever smoked pipes? [241]									
- Yes	281 (10.4%)	306 (11.4%)	53 (12%)	13 (12.7%)	14 (12.7%)	20 (12.1%)	268 (11.4%)	292 (11.3%)	33 (12%)
- No	2,168 (80.6%)	2,384 (88.6%)	387 (88%)	89 (87.3%)	96 (87.3%)	145 (87.9%)	2,079 (88.6%)	2,288 (88.7%)	242 (88%)

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
Have you ever been exposed to secondhand smoke? [71]	2,356 (87.6%)	2,426 (90.2%)	405 (92%)	92 (83.6%)	98 (89.1%)	153 (92.7%)	2,264 (87.8%)	2,328 (90.2%)	252 (91.6%)
If you have been exposed to secondhand smoke, specify by whom [134]									
- Spouse/ Partner	500 (18.6%)	524 (19.5%)	82 (18.6%)	15 (15.2%)	16 (14.5%)	33 (20%)	485 (19.7%)	508 (19.7%)	49 (17.8%)
- At Work	852 (31.7%)	919 (34.2%)	131 (29.8%)	34 (34.3%)	39 (35.5%)	43 (26.1%)	818 (33.3%)	880 (34.1%)	88 (32%)
- Home/ Work	85 (3.16%)	86 (3.2%)	11 (2.5%)	3 (3.03%)	3 (2.73%)	4 (2.42%)	82 (3.34%)	83 (3.22%)	7 (2.55%)
- Home/Leisure	17 (0.632%)	17 (0.632%)	3 (0.682%)	1 (1.01%)	1 (0.909%)	1 (0.606%)	16 (0.651%)	16 (0.62%)	2 (0.727%)
- Leisure	384 (14.3%)	408 (15.2%)	70 (15.9%)	14 (14.1%)	16 (14.5%)	27 (16.4%)	370 (15.1%)	392 (15.2%)	43 (15.6%)
- Leisure/ Work	102 (3.79%)	108 (4.01%)	26 (5.91%)	5 (5.05%)	6 (5.45%)	12 (7.27%)	97 (3.95%)	102 (3.95%)	14 (5.09%)

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
- Home/Leisure/ Work	112 (4.16%)	113 (4.2%)	23 (5.23%)	3 (3.03%)	4 (3.64%)	5 (3.03%)	109 (4.44%)	109 (4.22%)	18 (6.55%)
- Others at Home	249 (9.26%)	256 (9.52%)	53 (12%)	12 (12.1%)	13 (11.8%)	21 (12.7%)	237 (9.65%)	243 (9.42%)	32 (11.6%)
- Not exposed	255 (9.48%)	259 (9.63%)	41 (9.32%)	12 (12.1%)	12 (10.9%)	19 (11.5%)	243 (9.89%)	247 (9.57%)	22 (8%)
If you have been exposed to secondh and smoke, how many hours per day? [406]									
- <1	330 (12.3%)	406 (15.1%)	53 (12%)	7 (7.69%)	9 (8.18%)	10 (6.06%)	323 (14.7%)	397 (15.4%)	43 (15.6%)
- 2-6	1,038 (38.6%)	1,240 (46.1%)	218 (49.5%)	40 (44%)	50 (45.5%)	84 (50.9%)	998 (45.5%)	1,190 (46.1%)	134 (48.7%)
- >6	654 (24.3%)	778 (28.9%)	124 (28.2%)	32 (35.2%)	39 (35.5%)	50 (30.3%)	622 (28.4%)	739 (28.6%)	74 (26.9%)
- Not exposed	262 (9.74%)	266 (9.89%)	45 (10.2%)	12 (13.2%)	12 (10.9%)	21 (12.7%)	250 (11.4%)	254 (9.84%)	24 (8.73%)
Have you ever lived in a big	1,753 (65.2%)	1,849 (68.7%)	318 (72.3%)	73 (66.4%)	77 (70%)	131 (79.4%)	1,680 (65.1%)	1,772 (68.7%)	187 (68%)

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
city or near one for more than 10 years? [132]									
Have you ever worked with chemicals? [290]	304 (11.3%)	316 (11.7%)	47 (10.7%)	15 (13.6%)	15 (13.6%)	20 (12.1%)	289 (11.2%)	301 (11.7%)	27 (9.82%)
Have you ever been exposed to asbestos? [435]	124 (4.61%)	261 (9.7%)	42 (9.55%)	6 (5.45%)	12 (10.9%)	17 (10.3%)	118 (4.57%)	249 (9.65%)	25 (9.09%)
Have you ever been exposed to cadmium? [548]	11 (0.409%)	11 (0.409%)	2 (0.455%)	0 (0%)	0 (0%)	0 (0%)	11 (0.426%)	11 (0.426%)	2 (0.727%)
Have you ever been exposed to chromium? [525]	34 (1.26%)	34 (1.26%)	1 (0.227%)	0 (0%)	0 (0%)	0 (0%)	34 (1.32%)	34 (1.32%)	1 (0.364%)
Have you ever been exposed	6 (0.223%)	6 (0.223%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	6 (0.233%)	6 (0.233%)	0 (0%)

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
to berylliu m? [553]									
Have you ever been exposed to aluminu m? [520]	39 (1.45%)	39 (1.45%)	8 (1.82%)	3 (2.73%)	3 (2.73%)	5 (3.03%)	36 (1.4%)	36 (1.4%)	3 (1.09%)
Have you ever been exposed to silicon dust? [527]	32 (1.19%)	32 (1.19%)	3 (0.682%)	1 (0.909%)	1 (0.909%)	2 (1.21%)	31 (1.2%)	31 (1.2%)	1 (0.364%)
Have you ever been exposed to mixed sulfuric acid? [518]	41 (1.52%)	41 (1.52%)	7 (1.59%)	4 (3.64%)	4 (3.64%)	6 (3.64%)	37 (1.43%)	37 (1.43%)	1 (0.364%)
Have you ever been exposed to ether? [529]	30 (1.12%)	30 (1.12%)	4 (0.909%)	3 (2.73%)	3 (2.73%)	3 (1.82%)	27 (1.05%)	27 (1.05%)	1 (0.364%)
Have you ever been exposed	19 (0.706 %)	19 (0.706%)	2 (0.455%)	1 (0.909%)	1 (0.909%)	1 (0.606%)	18 (0.698%)	18 (0.698%)	1 (0.364%)

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
to coal? [540]									
Have you ever been exposed to nitrogen mustard? [555]	4 (0.149%)	4 (0.149%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	4 (0.155%)	4 (0.155%)	0 (0%)
Have you ever had a Pap smear? [905]									
- Last year	391 (14.5%)	554 (20.6%)	89 (20.2%)	12 (14%)	19 (17.3%)	25 (15.2%)	379 (22.3%)	535 (20.7%)	64 (23.3%)
- Last 5 years	282 (10.5%)	358 (13.3%)	61 (13.9%)	12 (14%)	13 (11.8%)	27 (16.4%)	270 (15.9%)	345 (13.4%)	34 (12.4%)
- No	1,112 (41.3%)	1,778 (66.1%)	290 (65.9%)	62 (72.1%)	78 (70.9%)	113 (68.5%)	1,050 (61.8%)	1,700 (65.9%)	177 (64.4%)
Have you ever had a mammography? [885]									
- Last year	533 (19.8%)	741 (27.5%)	115 (26.1%)	21 (23.9%)	30 (27.3%)	43 (26.1%)	512 (29.8%)	711 (27.6%)	72 (26.2%)
- Last 5 years	240 (8.92%)	295 (11%)	55 (12.5%)	15 (17%)	15 (13.6%)	21 (12.7%)	225 (13.1%)	280 (10.9%)	34 (12.4%)

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
- No	1,032 (38.4%)	1,654 (61.5%)	270 (61.4%)	52 (59.1%)	65 (59.1%)	101 (61.2%)	980 (57.1%)	1,589 (61.6%)	169 (61.5%)
Have you ever had a colonoscopy or sigmoidoscopy? [843]									
- Last year	264 (9.81%)	383 (14.2%)	68 (15.5%)	12 (13.5%)	15 (13.6%)	27 (16.4%)	252 (14.3%)	368 (14.3%)	41 (14.9%)
- Last 5 years	437 (16.2%)	692 (25.7%)	106 (24.1%)	18 (20.2%)	24 (21.8%)	46 (27.9%)	419 (23.8%)	668 (25.9%)	60 (21.8%)
- No	1,146 (42.6%)	1,615 (60%)	266 (60.5%)	59 (66.3%)	71 (64.5%)	92 (55.8%)	1,087 (61.8%)	1,544 (59.8%)	174 (63.3%)
Have you ever had a urological exam? [839]									
- Last year	311 (11.6%)	443 (16.5%)	73 (16.6%)	15 (16.9%)	18 (16.4%)	21 (12.7%)	296 (16.8%)	425 (16.5%)	52 (18.9%)
- Last 5 years	274 (10.2%)	371 (13.8%)	53 (12%)	10 (11.2%)	11 (10%)	21 (12.7%)	264 (15%)	360 (14%)	32 (11.6%)
- No	1,266 (47.1%)	1,876 (69.7%)	314 (71.4%)	64 (71.9%)	81 (73.6%)	123 (74.5%)	1,202 (68.2%)	1,795 (69.6%)	191 (69.5%)
Have you ever									

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
had a PSA test? [833]									
- Last year	638 (23.7%)	870 (32.3%)	138 (31.4%)	26 (28.9%)	32 (29.1%)	43 (26.1%)	612 (34.6%)	838 (32.5%)	95 (34.5%)
- Last 5 years	310 (11.5%)	402 (14.9%)	58 (13.2%)	13 (14.4%)	15 (13.6%)	27 (16.4%)	297 (16.8%)	387 (15%)	31 (11.3%)
- No	909 (33.8%)	1,418 (52.7%)	244 (55.5%)	51 (56.7%)	63 (57.3%)	95 (57.6%)	858 (48.6%)	1,355 (52.5%)	149 (54.2%)
Have you ever had a cardiological exam? [735]									
- Last year	651 (24.2%)	942 (35%)	154 (35%)	33 (36.3%)	39 (35.5%)	60 (36.4%)	618 (33.2%)	903 (35%)	94 (34.2%)
- Last 5 years	534 (19.9%)	762 (28.3%)	112 (25.5%)	20 (22%)	25 (22.7%)	41 (24.8%)	514 (27.6%)	737 (28.6%)	71 (25.8%)
- No	770 (28.6%)	986 (36.7%)	174 (39.5%)	38 (41.8%)	46 (41.8%)	64 (38.8%)	732 (39.3%)	940 (36.4%)	110 (40%)
Have you ever had a dermatological exam? [833]									
- Last year	312 (11.6%)	396 (14.7%)	73 (16.6%)	12 (13.6%)	12 (10.9%)	24 (14.5%)	300 (17%)	384 (14.9%)	49 (17.8%)

Original data: Variable [Missing]	**Original data: Total (2,690)**	**Imputed data: Total (2,690)**	**Balanced data: Total (440)**	**Original data: Lung cancer (110)**	**Imputed data: Lung cancer (110)**	**Balanced data: Lung cancer (165)**	**Original data: No cancer (2,580)**	**Imputed data: No cancer (2,580)**	**Balanced data: No cancer (275)**
- Last 5 years	343 (12.8%)	472 (17.5%)	85 (19.3%)	13 (14.8%)	16 (14.5%)	35 (21.2%)	330 (18.7%)	456 (17.7%)	50 (18.2%)
- No	1,202 (44.7%)	1,822 (67.7%)	282 (64.1%)	63 (71.6%)	82 (74.5%)	106 (64.2%)	1,139 (64.4%)	1,740 (67.4%)	176 (64%)
Do you often hear wheezing in your chest? [51]	441 (16.4%)	448 (16.7%)	104 (23.6%)	32 (29.1%)	32 (29.1%)	70 (42.4%)	409 (15.9%)	416 (16.1%)	34 (12.4%)
If you often hear wheezing in your chest, does it occur for several days or nights? [82]									
- Yes	296 (11%)	324 (12%)	67 (15.2%)	21 (19.8%)	23 (20.9%)	43 (26.1%)	275 (11%)	301 (11.7%)	24 (8.73%)
- No	121 (4.5%)	130 (4.83%)	25 (5.68%)	8 (7.55%)	9 (8.18%)	15 (9.09%)	113 (4.52%)	121 (4.69%)	10 (3.64%)
- No wheezing	2,191 (81.4%)	2,236 (83.1%)	348 (79.1%)	77 (72.6%)	78 (70.9%)	107 (64.8%)	2,114 (84.5%)	2,158 (83.6%)	241 (87.6%)
When wheezing occurs, do you also experience									

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
shortness of breath? [101]									
- Yes	103 (3.83%)	114 (4.24%)	24 (5.45%)	6 (5.83%)	8 (7.27%)	14 (8.48%)	97 (3.9%)	106 (4.11%)	10 (3.64%)
- No	295 (11%)	342 (12.7%)	70 (15.9%)	21 (20.4%)	25 (22.7%)	46 (27.9%)	274 (11%)	317 (12.3%)	24 (8.73%)
- No wheezing	2,191 (81.4%)	2,234 (83%)	346 (78.6%)	76 (73.8%)	77 (70%)	105 (63.6%)	2,115 (85.1%)	2,157 (83.6%)	241 (87.6%)
When you have wheezing, do you breathe normally between episodes? [189]									
- Yes	258 (9.59%)	397 (14.8%)	80 (18.2%)	12 (12.8%)	27 (24.5%)	50 (30.3%)	246 (10.2%)	370 (14.3%)	30 (10.9%)
- No	49 (1.82%)	55 (2.04%)	9 (2.05%)	5 (5.32%)	5 (4.55%)	5 (3.03%)	44 (1.83%)	50 (1.94%)	4 (1.45%)
- No wheezing	2,194 (81.6%)	2,238 (83.2%)	351 (79.8%)	77 (81.9%)	78 (70.9%)	110 (66.7%)	2,117 (88%)	2,160 (83.7%)	241 (87.6%)
In the past year, have you suffered from lung diseases	221 (8.22%)	230 (8.55%)	55 (12.5%)	17 (15.5%)	18 (16.4%)	30 (18.2%)	204 (7.91%)	212 (8.22%)	25 (9.09%)

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
that have limited your daily activities for more than a week? [85]									
If you have suffered from lung diseases that have limited your daily activities for more than a week in the past year, did you have an increased production of phlegm during such illnesses? [102]									
- Yes	178 (6.62%)	208 (7.73%)	45 (10.2%)	13 (12.5%)	17 (15.5%)	21 (12.7%)	165 (6.64%)	191 (7.4%)	24 (8.73%)
- No	46 (1.71%)	48 (1.78%)	12 (2.73%)	3 (2.88%)	3 (2.73%)	7 (4.24%)	43 (1.73%)	45 (1.74%)	5 (1.82%)

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
- No lung disease	2,364 (87.9%)	2,434 (90.5%)	383 (87%)	88 (84.6%)	90 (81.8%)	137 (83%)	2,276 (91.6%)	2,344 (90.9%)	246 (89.5%)
If you have suffered from lung diseases that have limited your daily activities for more than a week, have you had more than one illness of this kind in the past year? [119]									
- Yes	97 (3.61%)	114 (4.24%)	19 (4.32%)	5 (4.81%)	7 (6.36%)	10 (6.06%)	92 (3.73%)	107 (4.15%)	9 (3.27%)
- No	105 (3.9%)	134 (4.98%)	39 (8.86%)	10 (9.62%)	12 (10.9%)	20 (12.1%)	95 (3.85%)	122 (4.73%)	19 (6.91%)
- No lung disease	2,369 (88.1%)	2,442 (90.8%)	382 (86.8%)	89 (85.6%)	91 (82.7%)	135 (81.8%)	2,280 (92.4%)	2,351 (91.1%)	247 (89.8%)
Shortness of breath [155]									

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
- I stop because I struggle to breathe after 100 meters or after a few minutes of normal walking on flat ground.	27 (1%)	27 (1%)	4 (0.909%)	3 (2.91%)	3 (2.73%)	4 (2.42%)	24 (0.987%)	24 (0.93%)	0 (0%)
- I experience shortness of breath only when I walk quickly on flat ground or on a small incline.	546 (20.3%)	568 (21.1%)	86 (19.5%)	22 (21.4%)	24 (21.8%)	34 (20.6%)	524 (21.5%)	544 (21.1%)	52 (18.9%)
- I only experience shortness of breath from exertion.	1,871 (69.6%)	2,004 (74.5%)	335 (76.1%)	77 (74.8%)	82 (74.5%)	126 (76.4%)	1,794 (73.8%)	1,922 (74.5%)	209 (76%)
- No	91 (3.38%)	91 (3.38%)	15 (3.41%)	1 (0.971%)	1 (0.909%)	1 (0.606%)	90 (3.7%)	90 (3.49%)	14 (5.09%)

Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
Do you have a cough? [54]	1,202 (44.7%)	1,233 (45.8%)	243 (55.2%)	67 (60.9%)	68 (61.8%)	117 (70.9%)	1,135 (44%)	1,165 (45.2%)	126 (45.8%)
If you have a cough, is it daily? [419]									
- Yes	478 (17.8%)	587 (21.8%)	101 (23%)	27 (29%)	29 (26.4%)	45 (27.3%)	451 (20.7%)	558 (21.6%)	56 (20.4%)
- No	367 (13.6%)	652 (24.2%)	136 (30.9%)	26 (28%)	40 (36.4%)	62 (37.6%)	341 (15.7%)	612 (23.7%)	74 (26.9%)
- No cough	1,426 (53%)	1,451 (53.9%)	203 (46.1%)	40 (43%)	41 (37.3%)	58 (35.2%)	1,386 (63.6%)	1,410 (54.7%)	145 (52.7%)
If you have a cough, is it intermittent [465]									
- Yes	617 (22.9%)	800 (29.7%)	153 (34.8%)	34 (38.2%)	46 (41.8%)	67 (40.6%)	583 (27.3%)	754 (29.2%)	86 (31.3%)
- No	184 (6.84%)	440 (16.4%)	79 (18%)	14 (15.7%)	22 (20%)	38 (23%)	170 (7.96%)	418 (16.2%)	41 (14.9%)
- No cough	1,424 (52.9%)	1,450 (53.9%)	208 (47.3%)	41 (46.1%)	42 (38.2%)	60 (36.4%)	1,383 (64.7%)	1,408 (54.6%)	148 (53.8%)
Do you currently have phlegm? [63]	1,181 (43.9%)	1,214 (45.1%)	220 (50%)	53 (48.2%)	54 (49.1%)	100 (60.6%)	1,128 (43.7%)	1,160 (45%)	120 (43.6%)

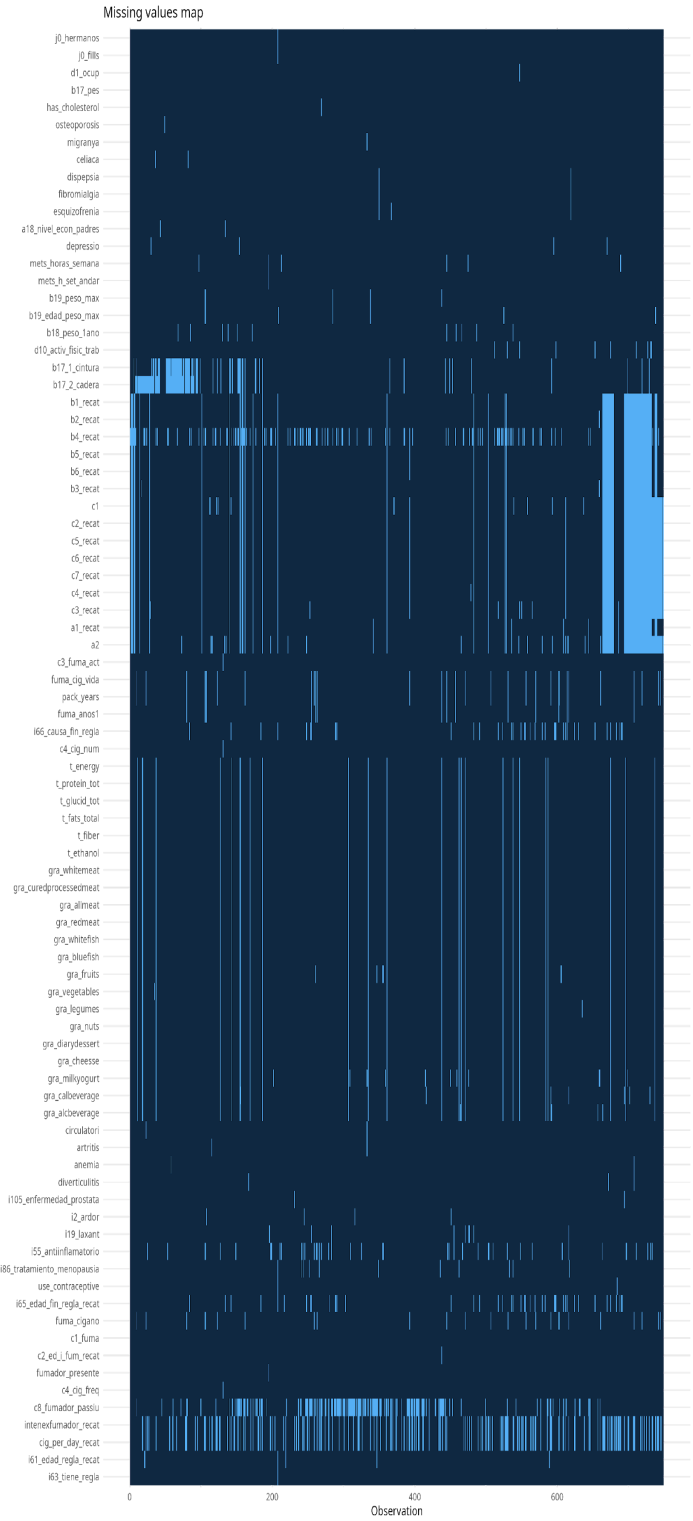
Original data: Variable [Missing]	Original data: Total (2,690)	Imputed data: Total (2,690)	Balanced data: Total (440)	Original data: Lung cancer (110)	Imputed data: Lung cancer (110)	Balanced data: Lung cancer (165)	Original data: No cancer (2,580)	Imputed data: No cancer (2,580)	Balanced data: No cancer (275)
If you have phlegm, is it mainly in the evening? [347]									
- Yes	161 (5.99%)	184 (6.84%)	36 (8.18%)	8 (8.16%)	8 (7.27%)	22 (13.3%)	153 (6.82%)	176 (6.82%)	14 (5.09%)
- No	749 (27.8%)	1,044 (38.8%)	174 (39.5%)	34 (34.7%)	46 (41.8%)	66 (40%)	715 (31.8%)	998 (38.7%)	108 (39.3%)
- No phlegm	1,433 (53.3%)	1,462 (54.3%)	230 (52.3%)	56 (57.1%)	56 (50.9%)	77 (46.7%)	1,377 (61.3%)	1,406 (54.5%)	153 (55.6%)
Peripheral oxygen saturation at rest (SpO2) (Mean/SD) [58]	97.1 (+-17.6)	97.1 (+-17.7)	96.7 (+-1.48)	96.5 (+-1.87)	96.5 (+-1.86)	96.6 (+-1.72)	97.1 (+-18)	97.1 (+-18.1)	96.7 (+-1.32)
Do you take bronchodilators to improve breathing? [502]	148 (5.5%)	164 (6.1%)	36 (8.18%)	11 (10%)	12 (10.9%)	25 (15.2%)	137 (5.31%)	152 (5.89%)	11 (4%)
Are you already being followed by a pulmonologist? [469]	137 (5.09%)	150 (5.58%)	35 (7.95%)	11 (10%)	11 (10%)	18 (10.9%)	126 (4.88%)	139 (5.39%)	17 (6.18%)

Original data: Variable [Missing]	**Original data: Total (2,690)**	**Imputed data: Total (2,690)**	**Balanced data: Total (440)**	**Original data: Lung cancer (110)**	**Imputed data: Lung cancer (110)**	**Balanced data: Lung cancer (165)**	**Original data: No cancer (2,580)**	**Imputed data: No cancer (2,580)**	**Balanced data: No cancer (275)**
Fagerstrom test score (Mean/ SD) [709]	4.69 (+-2.39)	4.07 (+-2.11)	4.26 (+-2.08)	5.02 (+-2.23)	4.39 (+-2.01)	4.42 (+-1.83)	4.68 (+-2.39)	4.05 (+-2.12)	4.17 (+-2.21)
Carbon monoxide level (Mean/ SD) [1,710]	2.72 (+-2.84)	2.72 (+-2.84)	3.01 (+-3.66)	3.36 (+-5.99)	3.36 (+-5.99)	3.34 (+-5.14)	2.68 (+-2.58)	2.68 (+-2.58)	2.79 (+-2.22)
Parts per million (ppm) of carbon monoxide (Mean/ SD) [1,726]	13.9 (+-10.2)	13.9 (+-10.2)	15.2 (+-9.94)	13.9 (+-9.88)	13.9 (+-9.88)	15 (+-9.55)	13.9 (+-10.2)	13.9 (+-10.2)	15.3 (+-10.2)
HADS Anxiety score (Mean/ SD) [297]	5.14 (+-3.65)	4.8 (+-3.47)	4.5 (+-3.32)	5.38 (+-3.99)	5.18 (+-3.84)	4.95 (+-3.43)	5.13 (+-3.64)	4.78 (+-3.45)	4.24 (+-3.23)
HADS Depression score (Mean/ SD) [170]	3.49 (+-2.96)	3.21 (+-2.77)	3.15 (+-2.68)	4.48 (+-3.44)	4.16 (+-3.31)	3.81 (+-2.93)	3.44 (+-2.93)	3.16 (+-2.74)	2.75 (+-2.43)
HADS Depression category [170]									
- Normal	2,260 (84%)	2,418 (89.9%)	405 (92%)	90 (85.7%)	94 (85.5%)	148 (89.7%)	2,170 (89.9%)	2,324 (90.1%)	257 (93.5%)

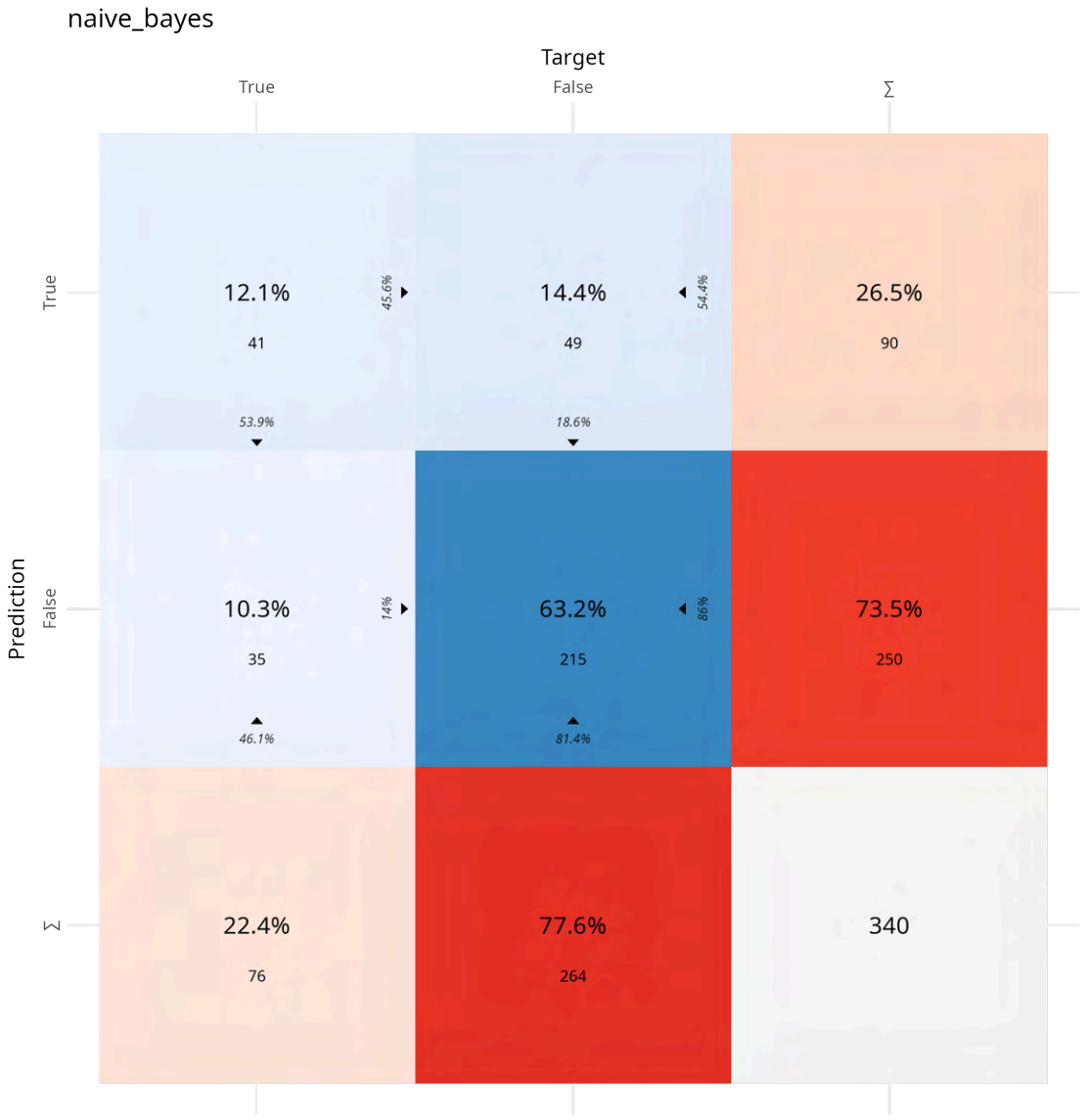
Original data: Variable [Missing]	**Original data: Total (2,690)**	**Imputed data: Total (2,690)**	**Balanced data: Total (440)**	**Original data: Lung cancer (110)**	**Imputed data: Lung cancer (110)**	**Balanced data: Lung cancer (165)**	**Original data: No cancer (2,580)**	**Imputed data: No cancer (2,580)**	**Balanced data: No cancer (275)**
- Borderline abnormal	187 (6.95%)	191 (7.1%)	25 (5.68%)	9 (8.57%)	9 (8.18%)	10 (6.06%)	178 (7.37%)	182 (7.05%)	15 (5.45%)
- Abnormal	73 (2.71%)	81 (3.01%)	10 (2.27%)	6 (5.71%)	7 (6.36%)	7 (4.24%)	67 (2.77%)	74 (2.87%)	3 (1.09%)
HADS Anxiety category [297]									
- Normal	1,823 (67.8%)	2,057 (76.5%)	346 (78.6%)	72 (75.8%)	81 (73.6%)	117 (70.9%)	1,751 (76.2%)	1,976 (76.6%)	229 (83.3%)
- Borderline abnormal	330 (12.3%)	362 (13.5%)	56 (12.7%)	11 (11.6%)	13 (11.8%)	31 (18.8%)	319 (13.9%)	349 (13.5%)	25 (9.09%)
- Abnormal	240 (8.92%)	271 (10.1%)	38 (8.64%)	12 (12.6%)	16 (14.5%)	17 (10.3%)	228 (9.92%)	255 (9.88%)	21 (7.64%)

7. Appendices (Figures)

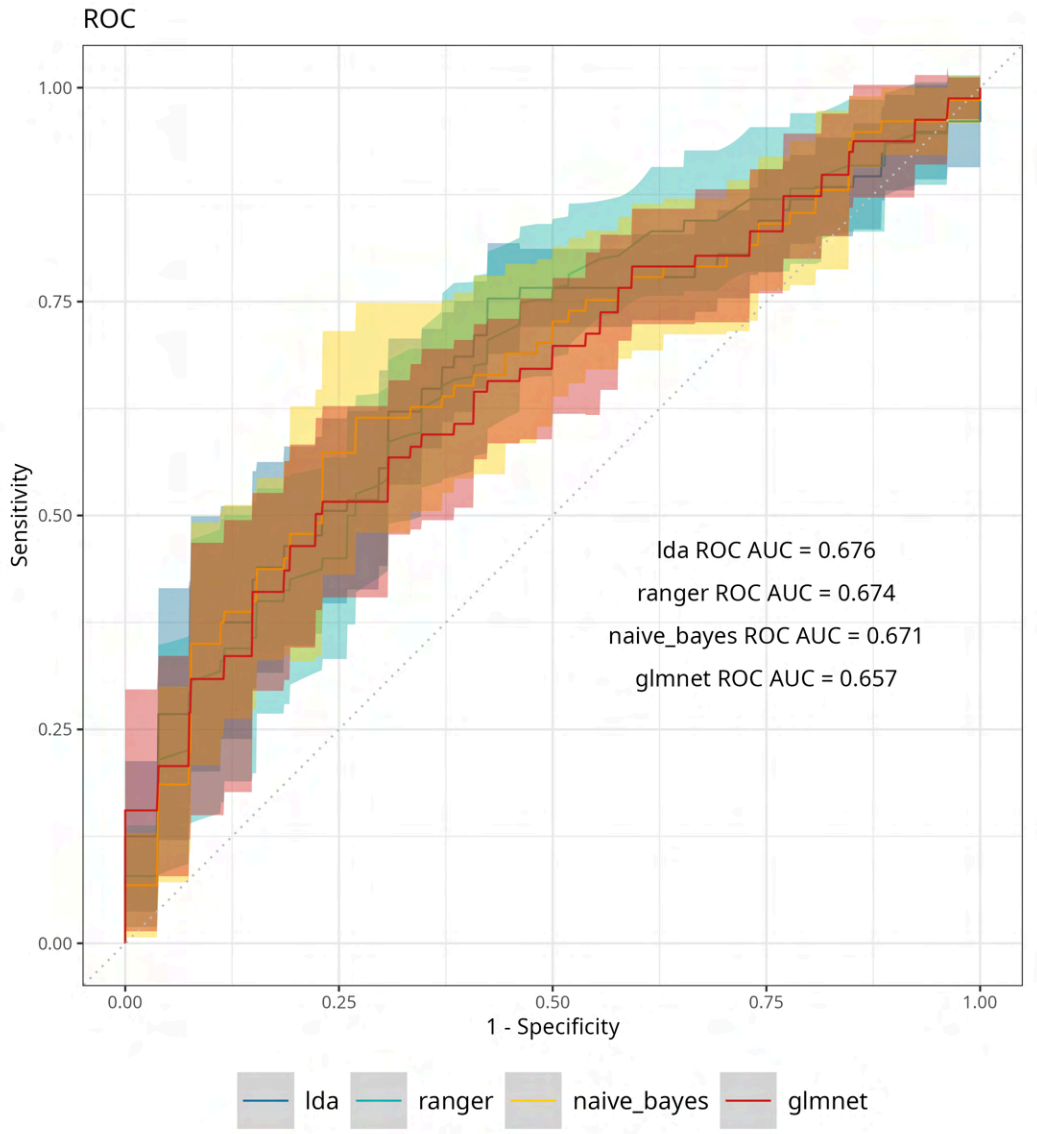
Supplemental Figure 1. Missing observations in the CRC screening study.



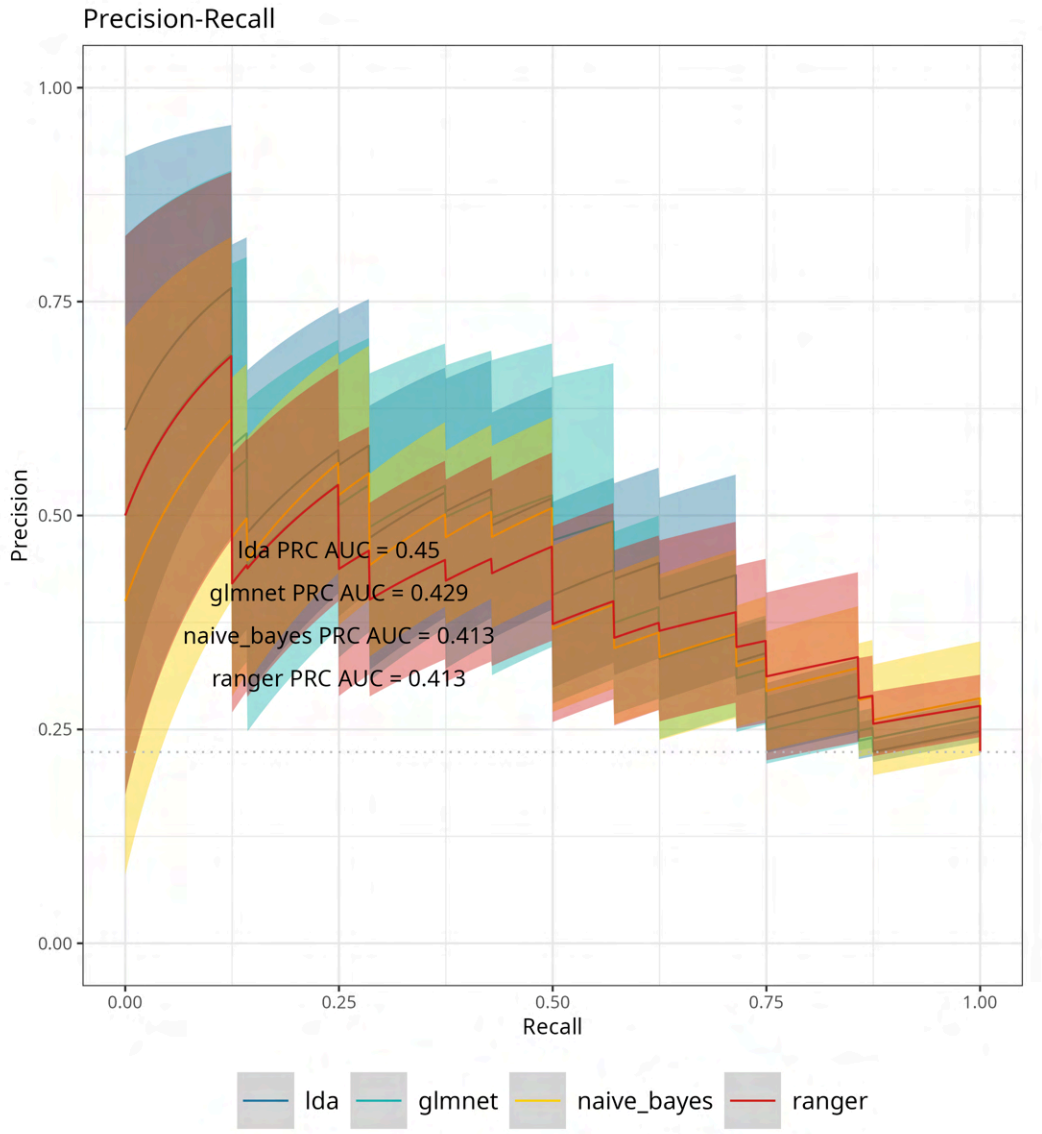
Supplemental Figure 2. Confusion matrix for the best performing model for the subgroup of females.



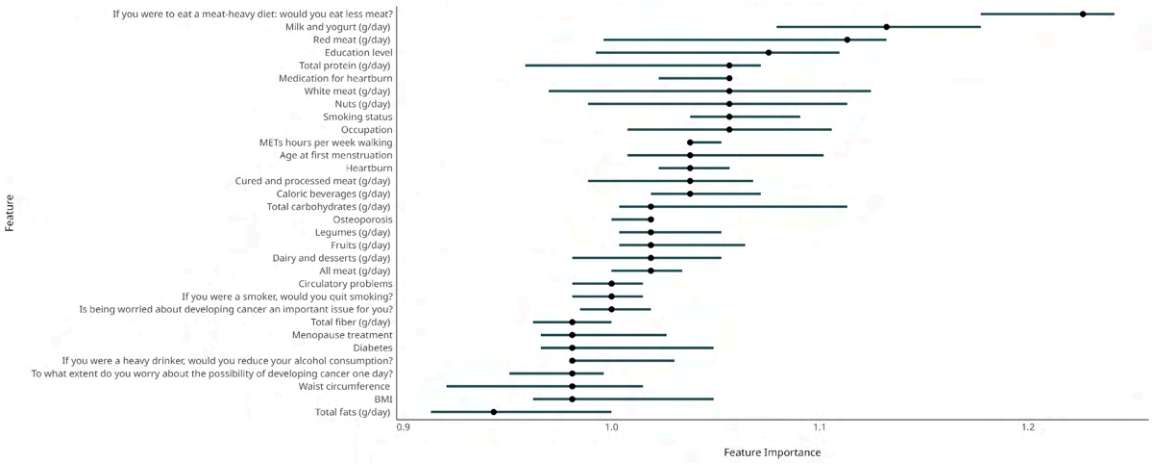
Supplemental Figure 3. Area Under the ROC curve for the subgroup of females.



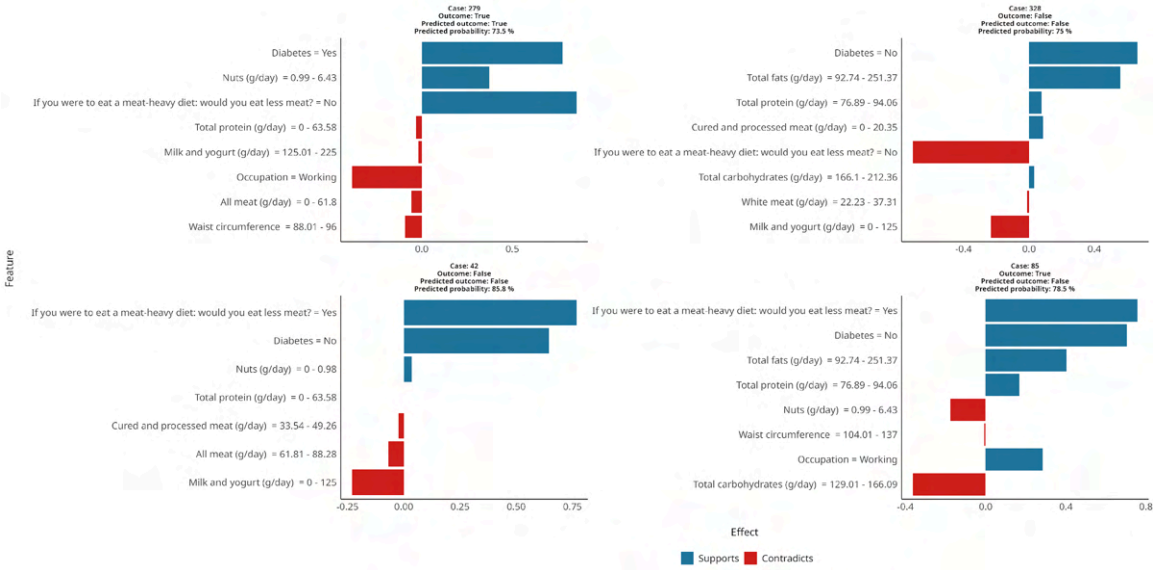
Supplemental Figure 4. Area Under the Precison-Recall curve for the subgroup of females.



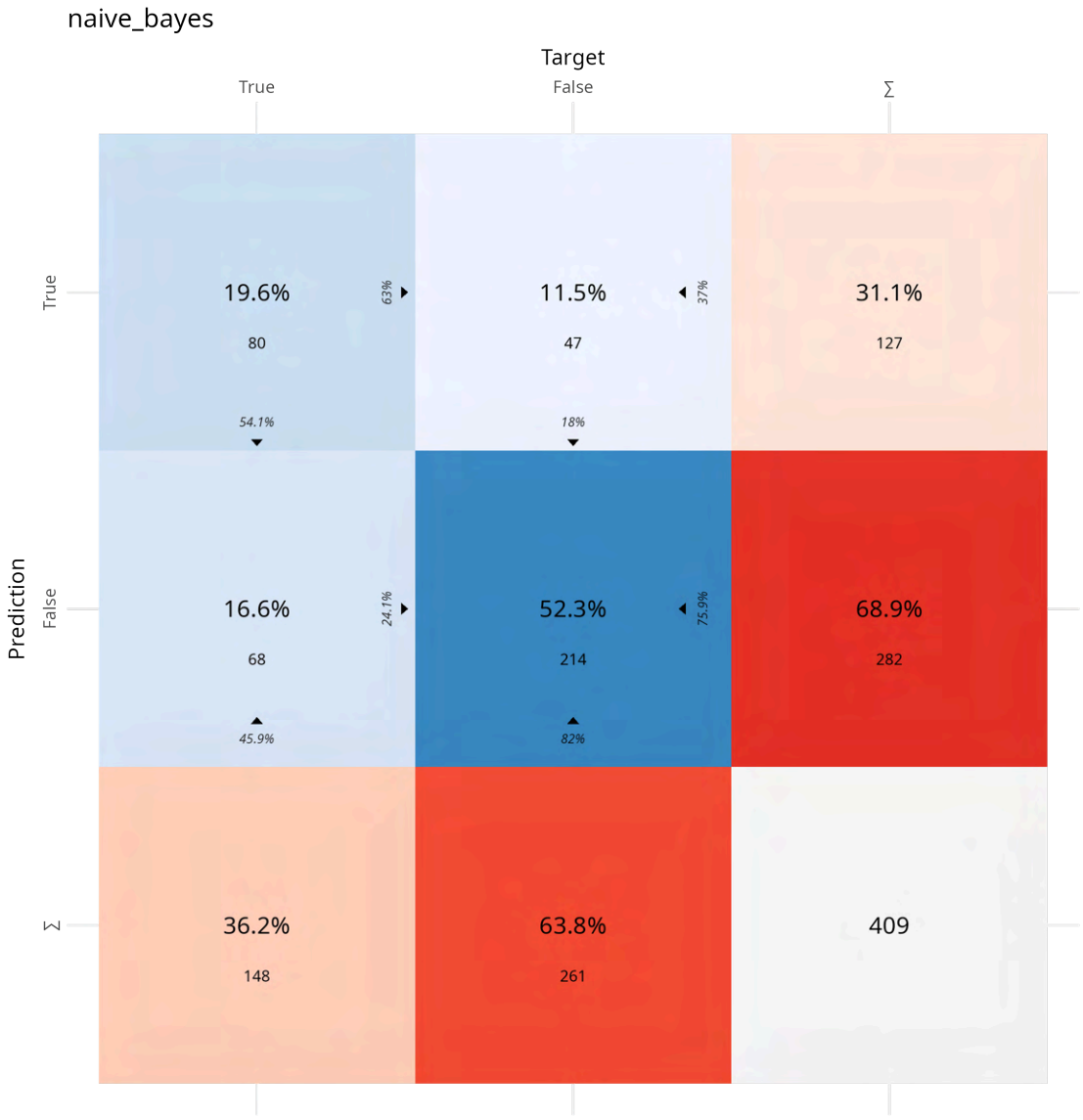
Supplemental Figure 5. Feature importance for the best performing model for the subgroup of females.



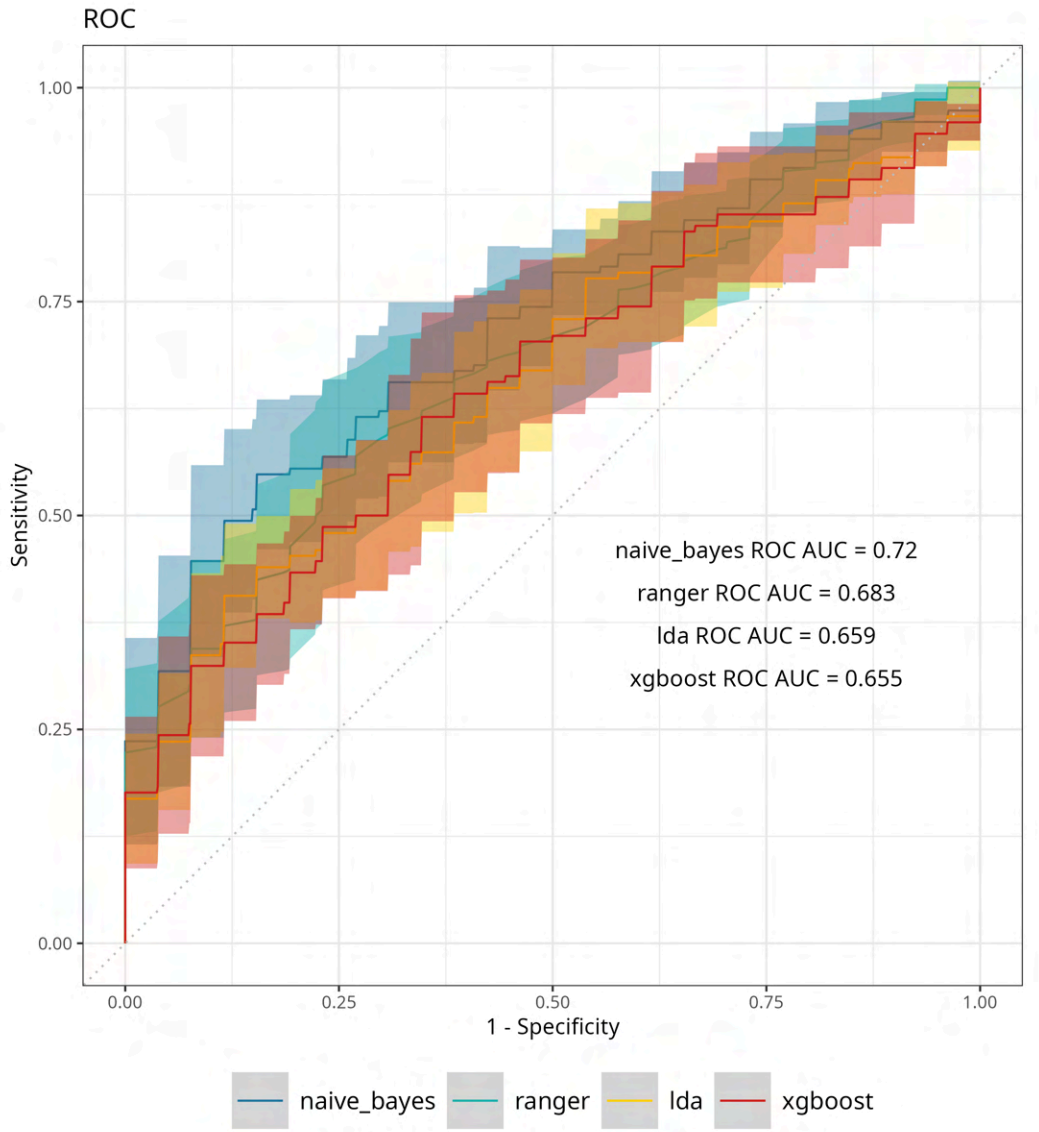
Supplemental Figure 6. LIME plot for the subgroup of females



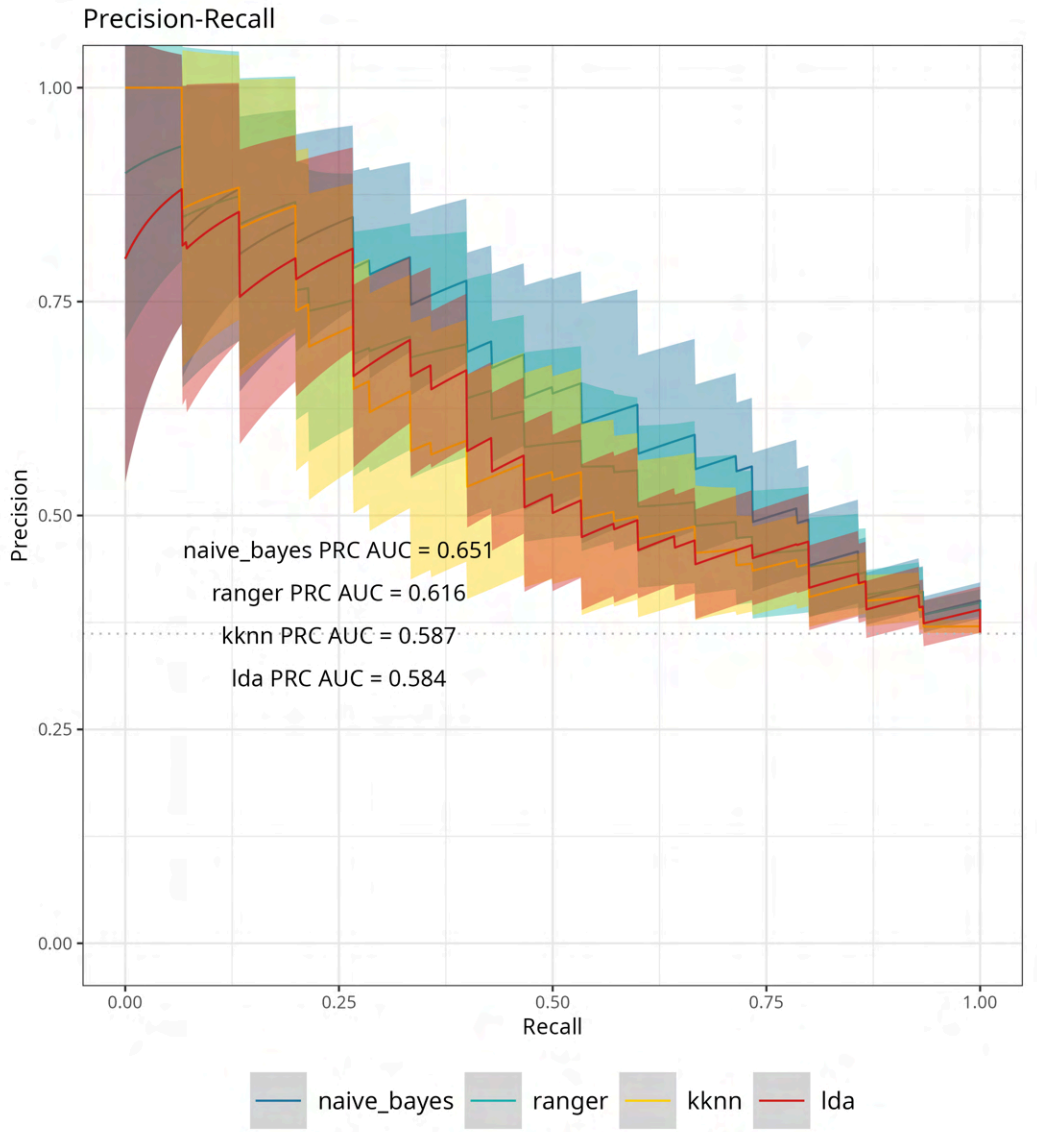
Supplemental Figure 7. Confusion matrix for the best performing model for the subgroup of males.



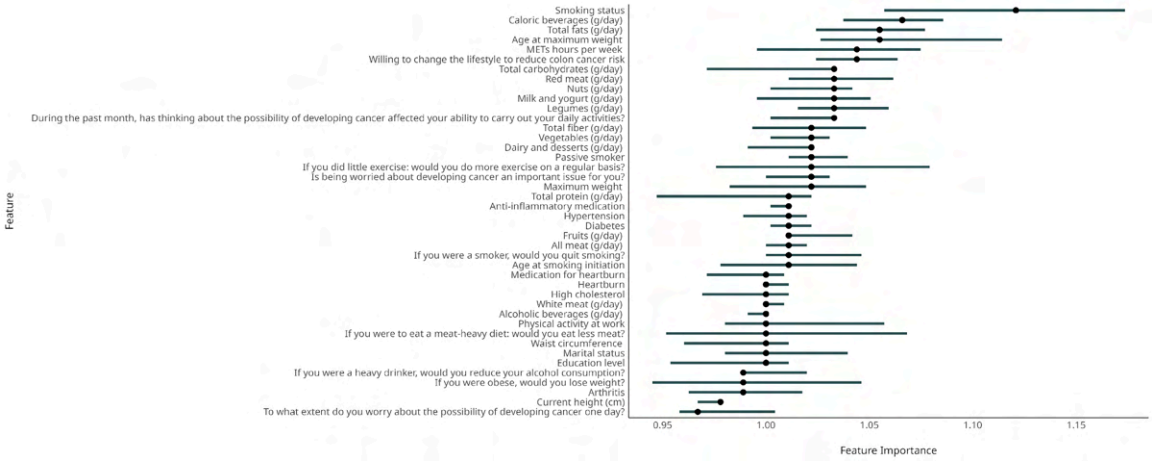
Supplemental Figure 8. Area Under the ROC curve for the subgroup of males.



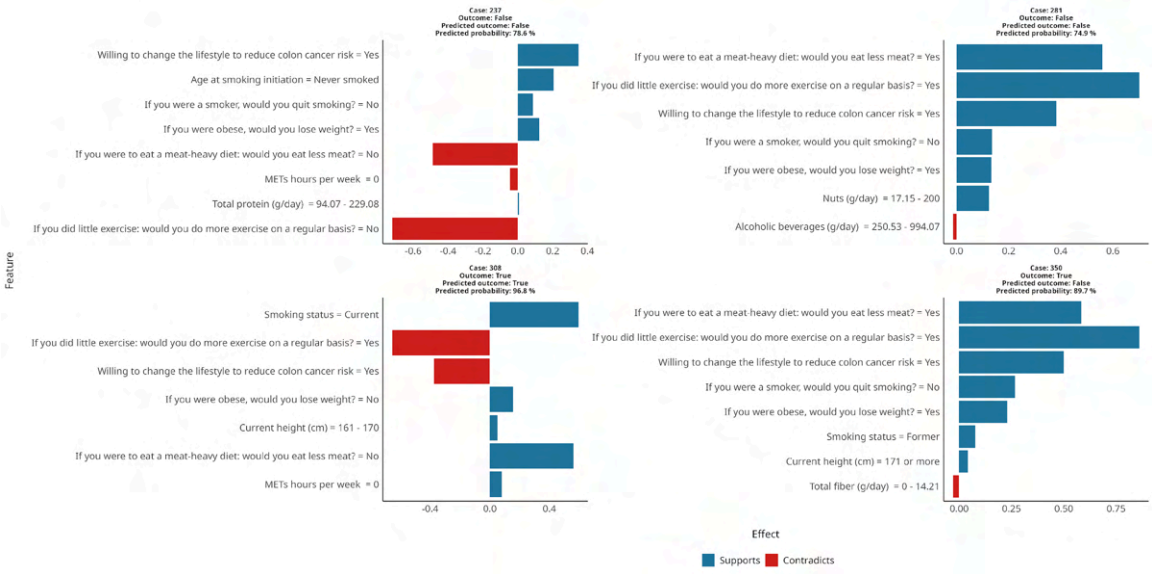
Supplemental Figure 9. Area Under the Precison-Recall curve for the subgroup of males



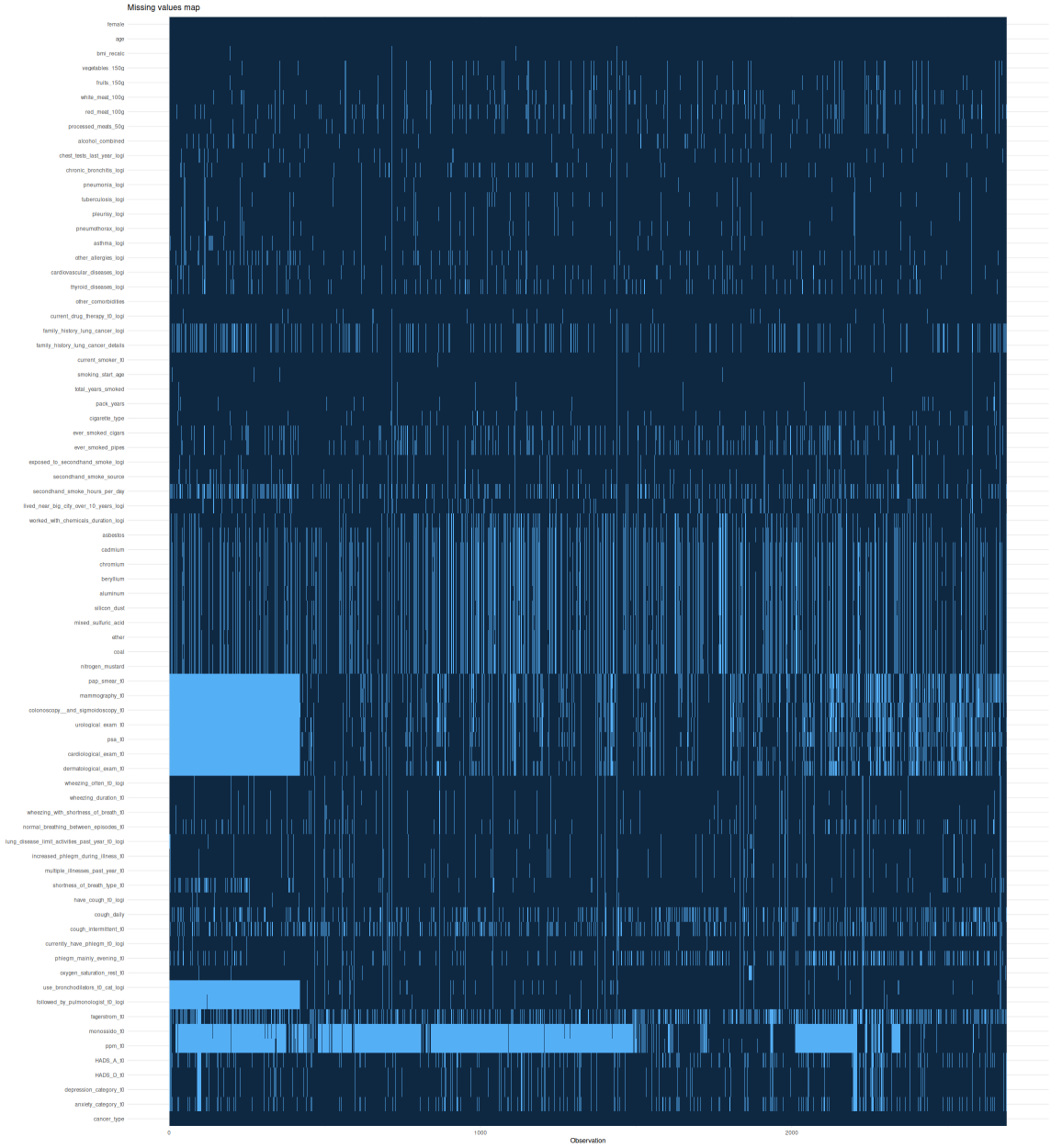
Supplemental Figure 10. Feature importance for the best performing model for the subgroup of males.



Supplemental Figure 11. LIME plot for the subgroup of males.



Supplemental Figure 12. Missing observations in the COSMOS study.


Missing values map
female
age
bmi_recalc
vegetables_150g
fruits_150g
white_meat_100g
red_meat_100g
processed_meats_50g
alcohol_combined
chest_tests_last_year_logi
chronic_bronchitis_logi
pneumonia_logi
tuberculosis_logi
pleurisy_logi
pneumothorax_logi
asthma_logi
other_allergies_logi
cardiovascular_diseases_logi
thyroid_diseases_logi
other_comorbidities
current_drug_therapy_t0_logi
family_history_lung_cancer_logi
family_history_lung_cancer_details
current_smoker_t0
smoking_start_age
total_years_smoked
pack_years
cigarette_type
ever_smoked_cigars
ever_smoked_pipes
exposed_to_secondhand_smoke_logi
secondhand_smoke_source
secondhand_smoke_hours_per_day
lived_near_big_city_over_10_years_logi
worked_with_chemicals_duration_logi
asbestos
cadmium
chromium
beryllium
aluminum
silicon_dust
mixed_sulfuric_acid
ether
coal
nitrogen_mustard
pap_smear_t0
mammography_t0
colonoscopy__and_sigmoidoscopy_t0
urological_exam_t0
psa_t0
cardiological_exam_t0
dermatological_exam_t0
wheezing_often_t0_logi
wheezing_duration_t0
wheezing_with_shortness_of_breath_t0
normal_breathing_between_episodes_t0
lung_disease_limit_activities_past_year_t0_logi
increased_phlegm_during_illness_t0
multiple_illnesses_past_year_t0
shortness_of_breath_type_t0
have_cough_t0_logi
cough_daily
cough_intermittent_t0
currently_have_phlegm_t0_logi
phlegm_mainly_evening_t0
oxygen_saturation_rest_t0
use_bronchodilators_t0_cat_logi
followed_by_pulmonologist_t0_logi
fagerstrom_t0
monoxsido_t0
ppm_t0
HADS_A_t0
HADS_D_t0
depression_category_t0
anxiety_category_t0
cancer_type
0
1000
2000
Observation

Version history

Version	Description	Date completed
v1.0	First version	19/05/2025
v2.0	Second version	25/07/2025
v2.1	Consortium revision	31/07/2025